Investigate Automatic Speech Recognition and Keyword Search for Very Low-Resource Language

Chongjia Ni
Institute for Infocomm Research (I2R), A*STAR, Singapore
e-mail: nicj@i2r.a-star.edu.sg

Bin Ma
Institute for Infocomm Research (I2R), A*STAR, Singapore
e-mail: mabin@i2r.a-star.edu.sg

Abstract—**In this paper, pronunciation lexicon, multi-lingual bottleneck features, semi-supervised learning, and data selection are investigated to help to improve the performance of automatic speech recognition (ASR) and keyword search (KWS) under very low-resource condition. For very low-resource condition, it is just about 3 hours of transcribed speech data, and there is no manual pronunciation for words in the transcription. According to our experiments on OpenKWS15 surprise language Swahili, some significant results can conclude. (1) Pronunciation lexicon has great influence on the performance of keyword search system at very limited language package (VLLP) condition when comparing with full language package (FLP) condition. (2) Multi-lingual bottleneck features (BNF) can improve the performance of ASR and KWS, and when combining with semi-supervised learning, the performance further improve. (3) Using large scale text corpus to train language model (LM), it can greatly improve the performance of KWS system and corresponding underlying ASR. When extending vocabulary size for keyword search, it can reduce out-of-vocabulary in keyword list, and thus slightly improve the performance of KWS system. (4) Initial transcription data selection is important to improve the performance of KWS and underlying ASR system.**

Keywords-multilingual bottleneck features; semi-supervised learning; pronunciation lexicon; data selection

I. INTRODUCTION

Low-resource or very low-resource speech and language technology has been a focus area for several research groups over the last several years. Due to resource constraints, automatic speech recognition (ASR) could not get state-of-the-art effect as rich-resource languages at low-resource or very low-resource condition. Several different technologies and approaches have been used for improving the performance of automatic speech recognition and keyword search (KWS), such as semi-supervised learning [1-5], multi-lingual bottleneck features [6-10]. When developing these technologies, different people or groups use different corpora, therefore, it is difficult to directly compare their effects for very low-resource keyword search and automatic speech recognition. In this paper, pronunciation lexicon, multi-lingual bottleneck features, semi-supervised learning, and data selection are investigated to help to improve the performance of ASR and KWS under very low-resource condition.

Based on NIST provided Swahili very limited language package (VLLP) and full language package (FLP), manual pronunciation lexicon and rule-based letter to sound (L2S) pronunciation lexicon have been compared for ASR and KWS. At VLLP condition, there is 5.4% word error rate augment when comparing L2S pronunciation lexicon based ASR with manual pronunciation lexicon based ASR. But at FLP condition, there is only 1.6% WER augment. Based on NIST provided Swahili VLLP data package and Swahili large-scale web text data, the effects of multilingual BNF, semi-supervised learning and "big" LM are evaluated. Only combining with multilingual BNF, there are about 3.4% WER reduction and 13.1% relative ATWV improvements. When combining with semi-supervised learning, the performance can improve further, and there is 3.1% WER reduction. Only using "big" LM to decode, it could get about 3.6~3.8% absolute WER reduction, and about 69.8~72.5% relative ATWV improvement.

More informative and representative utterances can help to build effective ASR system. Our proposed data selection approach [11] is used to select utterances for building ASR system. By combining with multilingual BNF, semi-supervised learning, and "big" LM, and there are 11.5% absolute WER reduction when comparing baseline system which combined monolingual BNF, semi-supervised learning and "big" LM, and there are 20% relative ATWV improvement. When comparing the baseline system without using multilingual BNF, semi-supervised learning and "big" LM, there are 15.3% absolute WER reduction and 106.9% relative ATWV improvement.

The rest of the paper is organized as follows. In Section 2, the related works are presented. In Section 3, the corpora used in our experiments are introduced. Experimental setup and experiments are presented in details at Section 4 and Section 5 respectively. In Section 6, we conclude the paper and give the future work.

II. RELATION TO PRIOR WORK

Pronunciation lexicon is a key component in current ASR system, and it is a bridge to connect the acoustic model and language model. In general, pronunciation lexicon is built manually by linguistic expert, and it takes a long time to build. For very low-resource or low-resource languages, it is difficult to find expert to build such lexicon. Grapheme based approaches [12-15] and data-driven lexicon discovery approaches [16-18] are proposed to automatically build pronunciation lexicon for different languages. In general, grapheme based ASR and KWS systems yield worse word error rate (WER) than phonetic lexicon based systems, but it

978-1-5386-0969-9/17/$31.00 ©2017 IEEE

is still interesting to see how such systems performance for KWS and ASR and their comparison with manual generated phonetic lexicon based system at very low-resource condition.

Borrowing rich-resource language feature extractor or acoustic model at deep neural network framework are common approaches for improving the performance of low-resource or very low-resource ASR or KWS system [6-10, 19]. The shared-hidden-layer multilingual deep neural network (SHL-MDNN) framework has been successfully applied in low-resource or very low-resource ASR or KWS [19]. According to previous study on Babel KWS data, cross-lingual BNF based knowledge transfer is better than cross-lingual DNN transfer [20]. It is also interesting to find the gain using the multilingual BNF to improve the KWS system performance in very low-resource condition.

Semi-supervised boosting trap approach [1-5] as a powerful technology can improve ASR system performance. At low-resource or very low-resource condition, there are different scales reduction in WER and augmentation in ATWV using semi-supervised learning. In general, confidence score based approaches are used to select utterances or utterance segments, and different decoding results can be fused to improve the quality of decoding transcription. By combining with "big" LM and lexicon contained large vocabulary, semi-supervised learning approach can get more effective results in terms of WER and ATWV according to our experimental results.

More informative and representative data can help to build effective ASR and KWS system. Our previous data selection approaches [11, 21-23] have indicated that development set matched submodular data selection approach could select more informative and respective data. By using these selected data and combining with multilingual BNF, semi-supervised learning, it can build more effective ASR and KWS system.

III. CORPORA DESCRIPTION

TABLE I. STATISTICS OF BABEL CORPORA

Language	Set	Data size (hours)	Lexicon size
Cantonese	FLP	69	18k
	LLP	10	5.9k
Pashto	FLP	72	21k
	LLP	10	7.0k
Turkish	FLP	68	46k
	LLP	10	12k
Tagalog	FLP	72	24k
	LLP	10	6.6k
Swahili	FLP	56	25k
	LLP	10	8k
	VLLP	3	5k

*after segmentation

The four languages (Cantonese, Pashto, Turkish, and Tagalog) of Babel corpora and OpenKWS15 surprise language Swahili are used in our ASR and KWS experiments. Each corpus mainly contains the conversational speech and less than 30% scripted speech. For the four languages, about 70 hours of data are recorded in each full language pack (FLP), and pronunciation lexicon only convers words appeared in training transcription. When using the four languages FLP to train multi-lingual bottleneck features extractor, conversational speech and scripted speech are both used. OpenKWS15 surprised language Swahili in very limited language pack (VLLP) contains about 3 hours of transcribed conversational speech data and about 85 hours of un-transcribed speech data. The 10 hours of development set Dev10h is used to evaluate system performance. Table 1 summarizes these corpora statistics.

IV. EXPERIMENTAL SETUP

A. Feature Extraction

1) Fbank and pitch related features

In this paper, fbank and pitch related features are used to train monolingual and multi-lingual BNF extractor. When extracting 22-dimensional fbank features, the high frequency is set to 3800. The MFCC and pitch related features are extracted and used to train initial GMMs, which used for force-alignment speech.

2) Monolingual and multi-lingual BNF

The shared-hidden-layer multi-lingual deep neural network (SHL-MDNN) [19] is used to train multi-lingual DNN feature extractor, in which the hidden layers are shared between different languages while the soft-max layers are language dependent. The SHL-MDNN contains 7 hidden layers, and except bottleneck layer, each hidden layer contains 2048 units. The bottleneck layer is 4th hidden layer, and it contains 42 units. For each language dependent soft-max layer, it contains about 4500 senones. In order to compare with multi-lingual BNF, the monolingual BNF also is extracted for Swahili. The monolingual BNF extractor contains 7 hidden layers, and each hidden layer contains 1024 units. The bottleneck layer is 4th hidden layer, and it also contains 42 units. The soft-max layer contains about 2000 senones. After exacting monolingual or multilingual BNFs, they are concatenated with fbank and pitch features, and the concatenated 117-dimensional features (fbank+pitch+ Δ+$\Delta\Delta$ + monolingual (multilingual) BNF) are used to train Swahili acoustic model.

B. Acoustic Modeling

The initial GMMs used for force-alignment were trained according to the maximum likelihood criterion. The LDA+MLLT+SAT transform is applied on the MFCC and pitch related features in a context window. The GMM systems based on LDA+MLLT+SAT features are used to force alignment speech and these alignments information are used for deep neural network training. At FLP condition, about 4500 senones are modeled, and At VLLP condition, about 2000 senones are modeled. The acoustic model used to speech recognition and keyword search is hybrid deep neural network, which is trained first using cross-entropy criterion and then using sMBR criterion.

C. *Language Modeling*

For the Babel project, one of biggest challenges is the sparse language model training data, especially at VLLP condition. Therefore, in order to reduce the degree of difficulty, NIST provides the Swahili text data to train LM for OpenKW15 evaluation. The original files, which are extracted from web pages or web queries, contain about 78m tokens. After data cleaning, there are about 44m tokens in these files. We use these cleaned files to train a general language model, and then interpolate with the language model trained using VLLP Swahili training transcription. In order to investigate the influence of different vocabulary size on the performance of ASR and KWS, top 100k, 150k, 200k and 250k vocabularies are selected and used to train different vocabulary size of language models.

D. *Speech Recognition and Keyword Search System*

All of speech recognition and keyword search experiments are carried out with the publicly available Kaldi toolkit [24]. When conducting keyword search experiments, we only search in-vocabulary queries in the word lattices using weighted finite state transducer approach provided by Kaldi. Through these keyword search experiments, the influence of different vocabulary sizes on keyword search can be found. The performance of keyword search systems is measured by ATWV [25]. WER is used to measure the performance of the underlying ASR systems. When conducting keyword system experiments, evaluation keyword list provided by NIST is used, which contains 4454 query items.

V. EXPERIMENTS

A. *Effect of Manual Pronunciation LEXICON*

In the first series of experiments, Swahili ASR and KWS systems are trained at VLLP condition using manual pronunciation lexicon and rule based L2S pronunciation lexicon respectively. The rule for generating L2S pronunciation lexicon just maps characters to phone according to documents provided by NIST. And in order to compare with ASR and KWS systems trained at VLLP condition, Swahili ASR and KWS systems at FLP condition are also built. Table 2 lists these experimental results. In Table 2, "Manual" column means that when building ASR and KWS systems, the manual pronunciation lexicon is used. "Rule" column means that when building ASR and KWS systems, the rule-based L2S pronunciation lexicon is used. When making these comparisons, the LMs are trained only using their training transcriptions.

TABLE II. WER AND ATWV COMPARISON FOR MANUAL PRONUNCIATION LEXICON AND RULE BASED L2S PRONUNCIATION LEXICON BASED SYSTEMS AT FLP AND VLLP CONDITION.

		FLP		VLLP	
		Manal	Rule	Manal	Rule
Dev10h	WER	41.6	43.2	65.7	71.1
	ATWV	0.5733	0.5398	0.2123	0.1882

From Table 2, we can find that at very low-resource condition, there is a great gap between manual pronunciation lexicon and rule based L2S pronunciation lexicon, and when there are more training data, the gap between them can reduce. In the following experiments (listed in 5.2~5.4), the rule based L2S pronunciation lexicon is used to build systems.

B. *Effect of Multilingual BNF*

Previous experiments on Vietnamese and Tamil languages have denoted that at LLP condition, multilingual BNF can improve the performance of KWS system and corresponding ASR system [20]. Fine-tuning the multilingual BNF on target data does not always improve the performance of KWS system. According to our experiments on Swahili VLLP system, there is no significant improvement on ATWV. Therefore, we just show the experimental results which there are no fine-tuning multilingual BNF on target data. Fig. 1 shows WER of different ASR systems. Fig. 2 shows ATWV of different KWS systems. In Fig. 1 and Fig. 2, "LM" means the LM is trained only using training transcription. "100k LM" means that we first use NIST provided web data to train LM, and then interpolate with LM trained with training transcription. When training LMs, top 100k vocabularies are selected and used. Others are similar.

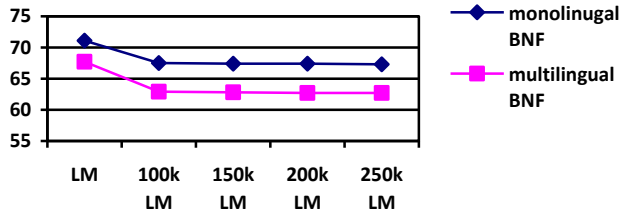

Figure 1. Performance comparison of different ASR systems

From Fig. 1, we can see that (1) When comparing monolingual BNF with multi-lingual BNF, multi-lingual BNF can provide better distinguish than monolingual BNF at very low-resource condition. There are about 3.4%~4.7% WER reductions. (2) It is almost no useful to extend vocabulary size from 100k to 250k for ASR when using same data to train LM.

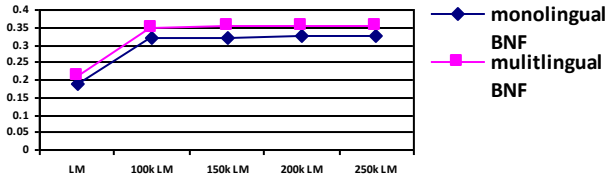

Figure 2. Performance comparison of different KWS systems

Fig. 2 shows the ATWV of different KWS systems at different LMs condition. As shown in Fig. 2, using the multilingual BNF can improve ATWV when comparing with using monolingual BNF. And using NIST provided text to train LM can improve ATWV greatly. It is just a little effect

to extend vocabulary size from 100k to 250k for improving ATWV.

C. Effect of Semi-Supervised Learning

Using multilingual BNF and NIST provided text trained LM, the untrancribed speeches are decoded, and then some speeches and corresponding decoding hypothesises are selected and used to re-train ASR system. In our experiments, only speech segments, which their confidence scores are greater than a threshold, are selected. In our experiments, the confidence threshold is 0.9. Table 3 shows these experimental results.

TABLE III. SEMI-SUPERVISED LEARNING EFFECTS FOR KWS AND ASR

		Trans. LM	Web LM interpolated Trans.LM			
			100k	150k	200k	250k
Supervised	WER	67.7	62.9	62.8	62.7	62.7
	ATWV	0.2129	0.3512	0.3524	0.3537	0.3566
Semi-supervised	WER	64.6	58.1	57.9	57.8	57.7
	ATWV	0.2257	0.3703	0.3763	0.3784	0.3798

In Table 3, "Tans. LM" means that the LM is trained only using 3 hours of Swahili VLLP training transcription. "Web LM interpolated Trans. LM" means that NIST provided web text is used to train different vocabulary size LMs, and then these LMs are interpolated with "Tans. LM" respectively.

From Table 3, we can see that using the semi-supervised learning approach, the performance of ASR system can improve, and there are about 3.1%~5.0% absolute WER reductions. For KWS systems, there are about 0.03 ATWV improvements.

D. Effect of initial training data

In section 5.1~5.3, the NIST provided VLLP data list is used to train all of listed ASR systems. In VLLP condition, training data used for building ASR system are important to help to get a better ASR system. In this section, we use our proposed submodular based data selection approach [11] to select same amount of speech data with NIST provided data to train ASR system. In order to select same amount of data, we first use VAD to extract voice part of data from NIST provided VLLP data, then get the total duration of voice part of data. When using our proposed submodular approach to select utterances for transcription, the total duration is used for optimal constraint. After building ASR system based on selected utterances, we also force alignment the train data in order to get the true voice part of data, then get the total duration of voice part, and compare the number with corresponding NIST provided VLLP data. There is almost no difference between them in voice part of data. Table 4 lists the experimental results.

In Table 4, "Active Learning" means that the KWS system and corresponding underlying ASR system are built using our selected 3 hours of transcribed data. "Active Leaning + multilingual BNF" means that when building the KWS and corresponding underlying ASR system, multilingual BNFs are concatenated with fbank and pitch related features, and the concatenated 117-dimensional features (fbank+pitch+ Δ+ΔΔ + multilingual BNF) are used to train acoustic model. "Active Leaning + multilingual BNF + Semi-supervised" means that semi-supervised learning is used and combined with "Active Leaning + multilingual BNF".

By selecting more informative and representative utterance, it can build more effective ASR and KWS system. When comparing with the system, which does not use "big" LM, multilingual BNF and semi-supervised, in Table 4 and Table 2, there are about 6.2% absolute WER reduction and 11.1% relative ATWV improvement. When comparing with system, which uses the "big" LM, multilingual BNF, and semi-supervised learning, in Table 4 and Table 3, there are about 2.0% absolute WER reduction and 2.5% relative ATWV improvement.

TABLE IV. ACTIVE LEARNING EFFECTS FOR KWS AND ASR

		Trans. LM	Web LM interpolated Trans.LM			
			100k	150k	200k	250k
Active Learning	WER	64.9	61.3	61.2	61.2	61.1
	ATWV	0.2091	0.3275	0.3286	0.3289	0.3298
Active Leaning + multilingual BNF	WER	61.9	57.9	57.7	57.6	57.6
	ATWV	0.2339	0.3638	0.3672	0.3674	0.3695
Active Leaning + multilingual BNF + Semi-supervised Learning	WER	60.2	56.1	56.0	55.9	55.8
	ATWV	0.2277	0.3825	0.3842	0.3885	0.3894

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we exploit the influence factors for improving the performance of ASR and KWS, which include pronunciation lexicon, multilingual BNF, semi-supervised learning, and initial transcribed data selection. Only from the view of improving the performance of KWS and ASR, "big" LM built using great scale text data is the primary factor, and it can remedy the shortcoming of acoustic model. Pronunciation lexicon has major influence on system performance in very low-resource condition, and this influence can make weak when providing more transcribed data to train ASR system. Multi-lingual BNF and semi-supervised learning both can be used to improve the performance of ASR and KWS, and it is about 10% absolute WER reduction and 17% relative ATWV improvement when comparing system without using multi-lingual BNF and semi-supervised learning at "big" LM condition. More informative and representative utterance can help to build more effective ASR and KWS system, and there are 2.0% absolute WER reduction and 2.5% relative ATWV improvement when comparing with the best system trained using NIST provided VLLP package and combining with multi-lingual BNF and semi-supervised learning technology at "big" LM condition.

Large scale text data not only can be used to improve LM, but also can be used for synthesis speech in order to augment speech for acoustic model training. In the future, we will exploit this approach to improve acoustic model for very low resource language.

REFERENCES

[1] F. Metze, A. Gandhe, Y. Miao, Z. Sheikh, Y. Wang, D. Xu, H. Zhang, J. Kim, I. Lane, W. K. Lee, S. Stuker, and M. Muler, "Semi-supervised Training in Low-resource ASR and KWS," in Proc. ICASSP 2015, pp.4699-4703.

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

[3] X. Cui, J. Huang, and J.-T. Chien, "Multi-view and Multi-objective Semi-supervised Learning for HMM-based Automatic Speech Recognition," IEEE Trans. on Audio, Speech, and Language Processing, 2012, 20(7):1923-1935.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition," in Proc. ICASSP 2002, pp. 845–848.

[13] E. Gunter Schukat-Talamazzini et al, "Automatic speech recognition without phonemes," in Proc. Eurospeech 1993.

[14] J. Dines and M. M. Doss. "A Study of Phoneme and Grapheme based Context-dependent ASR Systems," in Machine Learning for Multimodal Interaction. Springer, 2008, pp. 215–226.

[15] M. M. Doss et al. "Phoneme-grapheme based Speech Recognition System," in Proc. ASRU 2013, pp. 94-98.

[16] C.-y. Lee, and J. Glass, "A Nonparametric Bayesian Approach to Acoustic Model Discovery," in Proc. ACL 2012, pp.40-49.

[17] C.-y. Lee, Y. Zhang, and J. Glass, "Joint Learning of Phonetic Units and Word Pronunciations for ASR," in Proc. EMNLP 2013, pp. 182-192.

[18] L. Lu, A. Ghoshal, and S. Renals, "Acoustic Data-driven Pronunciation Lexicon for Large Vocabulary Speech Recognition," in Proc. ASRU 2013, pp.374-379.

[19] 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.

[20] 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.

[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.

[22] N. F. Chen, C. Ni, I-Fan Chen, S. Sivadas, V. T. Pham, H. Xu, X. Xiao, T. S. Lau, S. 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.

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

[24] 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.

[25] J. G. Fiscus, J.Ajot, J. S. Garofolo, and G. Doddingtion, "Results of the 2006 Spoken Term Detection Evaluation," in Proc. Interspeech 2007.