On the study of very low-resource language keyword search

Van Tung Pham*†, Haihua Xu†, Van Hai Do*†, Tze Yuang Chong†
Xiong Xiao†, Eng Siong Chng*† and Haizhou Li†‡
* School of Computer Engineering, Nanyang Technological University, Singapore
† Temasek Laboratories@NTU, Nanyang Technological University, Singapore
‡ Institute for Infocomm Research, Singapore

***Abstract*—In this paper we report our approaches to accomplishing the very limited resource keyword search (KWS) task in the NIST Open Keyword Search 2015 (OpenKWS15) Evaluation. We devised the methods, first, to attain better acoustic modeling, multilingual and semi-supervised acoustic model training as well as the examplar-based acoustic model training; second, to address the overwhelming out-of-vocabulary (OOV) KWS issue. Finally, we proposed a neural network (NN) framework to fuse diversified component systems, yielding improved combination results. Experimental results demonstrated the effectiveness of these approaches.**

Index Terms: speech recognition, low-resource, keyword search, multilingual training, semi-supervised training, system fusion

I. Introduction

Spoken Term Detection (STD) [1] or Keyword search (KWS) is a task of finding all occurrences of a keyword in a speech corpus. A common approach for the KWS task is to first use a speech recognizer to generate intermediate representations, e.g. n-best transcriptions or lattices, for spoken utterances; and then index them so that retrieval techniques can operate on such representations [2]–[8].

Building a keyword search (KWS) [9] system for a given language demands a lot of human-labeled data for training the automatic speech recognition (ASR) system. However, labeling the data is time-consuming and labor-intensive resulting in high development cost. This is particularly true for low-resource languages, uttered by a small number of speakers. Therefore, obtaining a decent KWS performance under very limited resource condition has been a challenge to the community.

Based on the above scenario, National Institute of Standards and Technology (NIST) has promoted low-resource language KWS evaluations in the past years, and the recent evaluation is the Open Keyword Search 2015 (OpenKWS15) [1]. In this evaluation, NIST has formulated a very limited language pack (VLLP) condition, in which only about 3 hours of transcribed speech can be used to build the ASR system. Such challenge motivates researchers to examine various techniques such as multilingual or semi-supervised training to improve the performance of the acoustic modeling.

[1] http://www.nist.gov/itl/iad/mig/openkws15.cfm

Multilingual training, particularly for the deep neural network (DNN) based multilingual training approach, has previously been proved to improve the ASR performance for the low-resource languages. This is because the DNN can be understood as a cascade of a feature extractor and a classifier. That is, the lower hidden layers are mainly responsible for feature extraction, while the soft-max layer is for classification. If we train the hidden-layers by multilingual data to extract some shared cross-language features, and keep the soft-max layer language-dependent, we are thus able to benefit from multilingual data to prepare the recognizer for a low-resource target language.

Semi-supervised training is an alternative for acoustic modeling in low-resource data scenario [10]–[12]. In semi-supervised training, people try to use ASR results as the reference to make use of unlabeled data. However, since the seed model is always built on a very limited transcribed data, the performance of such model remains much to be desired. Note that in semi-supervised training, the machine transcribed data are imperfect, hence data selection becomes critical.

In this paper, our aim is to build both the state-of-the-art ASR and the KWS systems under the VLLP condition. To build better ASR system, various techniques are employed. First, we used 6 languages from previous evaluations, i.e. OpenKWS13 [13] and OpenKWS14 [14], as rich-resource languages to build the multilingual DNN systems using the shared-hidden layer training method [15]. We demonstrated their effectiveness in terms of both the multilingual bottleneck feature (MBNF) and the multilingual DNN-HMM hybrid systems. Second, we took the multilingual training boosted ASR systems as seed models to conduct semi-supervised training, yielding further improved results. Finally, we also proposed an examplar based acoustic modeling approach, using the MBNF as front-end feature, and its efficiency under the VLLP condition was demonstrated.

Since better ASR system does not necessarily lead to better KWS performance, we aim to improve the KWS results from two aspects: 1) we attempt to account for the out-of-vocabulary (OOV) KWS problems through an unsupervised subword modeling approach; 2) we study a system fusion method using a neural network (NN) classifier to achieve an improved fusion result.

This paper is organized as follows. In Section II, we briefly

TABLE I
THE DISTRIBUTION OF THE SWAHILI ACOUSTIC DATA IN THE VLLP CONDITION

Acoustic Data	Speech data(hour)
Training	3.15
Tuning	3.16
Development	10.65
Unlabeled	87.79

TABLE II
THE DISTRIBUTION OF THE MULTILINGUAL TRAINING DATA

Full Language Pack (FLP)	Speech data (hour)	Lexicon Size (k)
Cantonese	141.3	20
Turkish	77.2	45
Pashto	78.4	23
Tagalog	84.5	33
Vietnamese	87.7	9.6
Tamil	69.3	69

describle the OpenKWS task. In Section III, we present our approaches to improving the ASR performance under low-resource condition. Section IV describes our strategies, i.e. subword modeling and NN fusion, to improving the KWS results. Experimental results are presented in Section V. Finally, Section VI concluses this paper.

II. NIST OPENKWS15 TASK

A. ASR data description

In the NIST OpenKWS15 evaluation for the VLLP task, participants are given a surprise language that is unknown until the evaluation date. The surprise language is Swahili this year. The released data include four data sets, namely Training, Tuning, Development, and Unlabeled data: 1) The acoustic data are collected from various real noisy scenes and telephony conditions. Three hour data is transcribed for training. Besides, there are 3 hour of tuning data and 10 hour of development data. 2) No lexicon is provided. 3) Text data for language modeling is provided by the organizer. The data is collected from various public available websites. Table I summarizes the details of the acoustic data. We note that the tuning data is only used for the ASR system tuning.

The text data contains 84M words altogether. It is used to establish the lexicon of 350K words in size, and to build trigram language models. Since Swahili is an agglutinative language, new words and long words are common, leading to high OOV rate for a given vocabulary. For instance, even with the 350K vocabulary lexicon, the OOV rate on the *dev* data is still 7.4%. Besides, the pronunciation of each word is represented as a grapheme string [16]. This is because no lexicon expertise knowledge is available.

To overcome the data limitation, we use the unlabeled data, as described in Table I, during a semi-supervised training. In addition, multilingual training is another intended approach to making improvement. The detailed distribution of the multilingual data, which is provided by NIST, is presented in Table II.

TABLE III
THE SUMMARY OF THE KEYWORD LISTS

Statistics	*dev*	*eval*
#KW	2480	4464
KW OOV rate	12.02%	7.07%
#word per keyword	1.54	1.72

B. KWS data description

In addition to the data for training ASR systems, NIST also provided two sets of keyword (KW) lists to evaluate the performance of the KWS system. One is the development set, and the other is the evaluation set. In table III, we summarize the development (dev) and evaluation ($eval$) keyword lists

C. Term weighted value metric

To evaluate the KWS performance, NIST defines the term-weighted value (TWV) which integrates the miss rate and false alarm rate of each keyword into a single metric [1]. Specifically for a keyword q_k we have:

$$TWV(q_k,\theta) = 1 - \frac{1}{M_k}\sum_{k=1}^{M_k}((P_{miss}(q_k,\theta) + \beta P_{fa}(q_k,\theta)) \quad (1)$$

where $P_{miss}(q_k,\theta)$ and $P_{fa}(q_k,\theta)$ are the probability of miss and the probability of false alarm of the query q_k respectively with respect to a detection threshold θ. The weight β is related with the prior probability of a keyword, and the cost ratio between the false alarm and the miss errors.

Actual term-weighted value (ATWV) is the average TWV of all keywords at a chosen decision threshold. The Detection Error Tradeoff (DET) curve is another evaluation metric for KWS performance [1]. To give an overview of the system performance, a DET curve visualized the overall performance of a STD system by plotting the tradeoff between probability of miss P_{miss} versus probability of false alarm P_{fa}. In this paper, both ATWV and DET curves are used for performance evaluation and all results are reported on the development data.

III. APPROACHES TO IMPROVING THE ASR PERFORMANCE UNDER LOW-RESOURCE CONDITION

In this section, we describe our three approaches to improving the performance of the ASR systems under the low-resource condition. They are multilingual training, semi-supervised training and examplar-based acoustic modeling.

A. Multilingual training

In this work, we examine two recipes to exploit the multilingual training to make improvement on the low-resource speech recognition as mentioned previously. One is to use the multilingual DNN (MDNN) to do DNN-HMM acoustic modeling directly. The multilingual DNN training framework is the same with [15], [17]. It is illustrated in Figure 1.

Once we finish the MDNN training, we change the softmax layer, using the target language. Consequently, MDNN based cross-lingual transfer is realized by tuning the MDNN with limited data. Normally cross-entropy training can yield

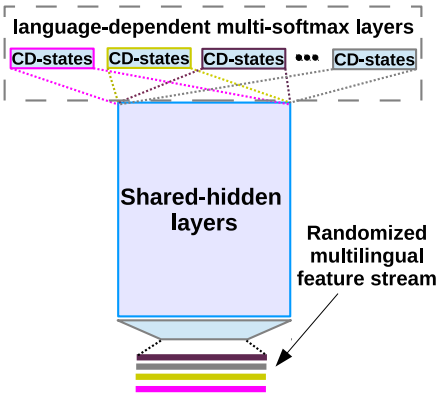

Fig. 1. *Illustration of the Multilingual DNN training using shared-hidden layer framework*

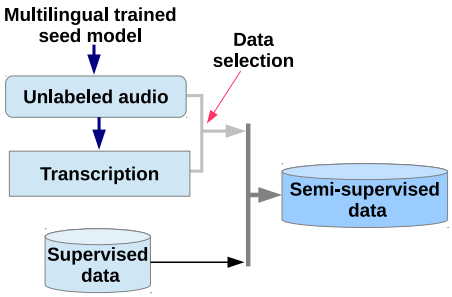

Fig. 2. *Illustration of the procedure of the semi-supervised data generation*

improved results, but further improvement can be obtained by the sequential training [18].

The other approach is to use the multilingual training data to train a multilingual bottle-neck (MBN) DNN as a feature extractor. The framework is similar to Figure 1, except that a hierarchical bottle-neck structure [10], [11], [19] is used.

In all cases, we used 22 dimensional filter-bank features, plus 3 dimensional pitch features [20]. For the MDNN training, the window size is 21 frames; for the MBN DNN training, the window size is 11 frames. Both delta and double-delta features are appended. Mean and variance normalization are applied. For details of our multilingual training, readers can refer to [15].

B. Semi-supervised training

As mentioned in Section I, two factors are critical to the success of the semi-supervised training. One is to have a good seed model, and the other is to have an effective data selection method. In our work, the seed model was obtained from the multilingual training. For the data selection, a word-segment based data selection method was employed, using the confidence score estimated with Minimum Bayes Risk decoding method [9]. The diagram of the semi-supervised data generation is illustrated in Figure 2

After the semi-supervised data generation is done, the semi-supervised training is conducted. The main steps are: 1) retrain the HMM-GMM system appropriately; 2) redo cross-lingual training on the multilingual DNNs using the new tied-states from step 1); 3) fine-tune the DNNs with the supervised data only.

C. Examplar-based acoustic model

Recently, the kernel density model [21] - a special case of the exemplar-based approach [22] was applied for acoustic modeling for low-resource languages. Unlike the parametric models, the kernel density model is a non-parametric technique that uses the training samples directly without estimating model parameters. This allows us to make full use of the limited training data. To the best of our knowledge, this approach has not been studied in KWS applications.

In this approach, instead of using a GMM to model the feature distribution of a triphone tied-state as in the conventional HMM/GMM acoustic model, we use the kernel density model similar to the one used in [21], [23]. Specifically, the likelihood of feature vector $\boldsymbol{o}_t$ for speech class i.e. HMM tied-state s_j, is estimated as follows:

$$p(\boldsymbol{o}_t|s_j) = \frac{1}{ZN_j}\sum_{i=1}^{N_j}\exp\Big(-\frac{||\boldsymbol{o}_t-\boldsymbol{e}_{ij}||^2}{\sigma}\Big) \quad (2)$$

where $\boldsymbol{e}_{ij}$ is the i^{th} training exemplar of class s_j, $||\boldsymbol{o}_t-\boldsymbol{e}_{ij}||$ is the Euclidean distance between $\boldsymbol{o}_t$ and $\boldsymbol{e}_{ij}$, σ is a scale variable, N_j is the number of exemplars in class s_j, and Z is a normalization term to make Eq. (2) be a valid distribution.

From Eq. (2), the likelihood function is mathematically similar to a GMM with a shared scalar variance for all dimensions and Gaussians. Effectively, Eq. (2) puts a Gaussian-shaped function at each training exemplar and sums all these Gaussians with a normalization factor to the likelihood function. In Section V-A2 and Section V-B1, we will demonstrate its advantages over the conventional GMM-HMM system for the speech recognition and the KWS tasks respectively.

IV. Approaches to Improving the KWS Performance

Given the improved ASR systems, our next objective is to achieve the best KWS performance for each single system, and to fuse those single systems appropriately. One of the major challenges with the KWS system is how to deal with the OOV problem, which is particularly pronounced for those agglutinative languages. In this work, we employ the subword modeling method to alleviate the effect of the problem. Furthermore, instead of using the conventional system fusion method, we propose a neural network based system fusion method. In what follows, we demonstrate the contribution of these two methods on improving the KWS performance.

A. Subword modeling

A wide range of subword units have been proposed for the KWS task, including linguistical units such as phones [5], [24], [25], syllables [26]–[28], as well as data driven units called morpheme [29], [30], word-fragment [31], multigrams [7], graphone [32] and particles [33]. It requires expert knowledge to establish a set of linguistic units for a target language, which

is not available in our case. As a result, we choose a data driven subword units, i.e. morpheme, for our KWS systems.

We adopt the unsupervised method to train word segmentation model using Morfessor [34]. Specifically, we first collect the word-count list to train a segmentation model. Then such model is used to decompose our word lexicon into a morpheme lexicon. Since grapheme-based pronunciations are used in all our experiments, it is straightforward to generate the pronunciations for each morpheme. Accordingly, we also used the model to decompose word-based text into morpheme-based text to build morpheme language models. Note that as morpheme and word systems share the same phone set, acoustic model built from the word system is still applicable to the subword systems.

B. Classifier-based approach for system fusion

It is common that we develop several KWS of different architectures and combine them together to form a mixture-of-expert. The way to combine multiple KWS system outputs is called system fusion. The first step is to align the detections returned by all KWS systems to find groups of overlapped detections, denoted as h_i where i = 1 ... n, then we merge h_i into a final detection h. Determining the score of the final detection h is the key issue in view of the variety of different systems and keywords.

One can use a arithmetic function, such as CombSum or WCombMNZ as introduced in [35], to estimate the score of h. However, such rule-based functions are ignorant of the relationship between the input information and the combined score. We propose to use a discriminative classifier, i.e. neural network for the fusion task. In this approach, for each combined detection h, scores of h_i as well as other system and keyword characteristics are taken into consideration to train the binary classifier. The outputs label 0/1 for the classifier correspond to the false alarm/correct status of the detection h. This approach provides a uniform framework to incorporate various features into the fusion. This work is similar to our previous study [36] except the feature sets used for fusion. Specifically, for each detection h_i from the i^{th} system, the following features are extracted:

- Word posterior probability calculated from the ASR lattice output
- The corresponding Keyword Specific Threshold (KST) score and KST decision, estimated as in [37]
- The corresponding Sum-To-One (STO) score [35] and β-STO [38]
- The PFA score, and PFA-KST score estimated as in [39]
- The rank of h_i in the detections lists of the i^{th} system
- The indicator value to indicate that h_i is null or not

In the case that the i^{th} system does not have h_i, i.e. h_i is null, those features are set to default values (0 for all features except the rank feature. For such a feature, since higher rank denotes less confidence, the default value is 20000).

We also collect a list of global features that capture the characteristic of the keyword and the detection h:

- Duration of the detection in seconds
- Number of words of the keyword
- Number of vowel and consonant letters of the keyword
- Speaking rate (i.e. duration/ number of letters)
- Location of the detection h in the utterence (at begining or middle or at the end).

TABLE IV
THE EFFECTIVENESS OF THE MULTILINGUAL AND SEMI-SUPERVISED TRAINING, TAKING THE MONO-LINGUAL TRAINING AS BASELINE

System	WER (%)
Monolingual DNN-HMM hybrid	67.9
Multilingual DNN-HMM hybrid	59.7
+ semi-supervised training	55.4
MBNF based DNN-HMM hybrid	56.5
+ semi-supervised training	54.6

TABLE V
THE EFFECTIVENESS OF THE EXAMPLAR BASED ACOUSTIC MODELING

System	WER (%)
MBNF based DNN-HMM hybrid	56.5
MBNF based examplar system	54.9

V. EXPERIMENTAL RESULTS

In this section, we report our experimental results in OpenKWS15 evaluation. We first report our ASR system performances with the various techniques as mentioned in section III, then we report our KWS performances of individual system as well as fusion methods.

A. Speech recognition performance

1) Multilingual and semi-supervised training: Table IV reveals the effectiveness of the multilingual and semi-supervised training. The results suggest that: 1) The multilingual training is very effective to improve the system performance under the very low-resource (VLLP) condition; it makes 12.0% relative WER reduction (WERR) over the monolingual baseline (from 67.9% down to 59.7%). 2) Semi-supervised training is still important to make further improvement. In our two cases, it realizes 7.2% and 3.4% relative WERRs respectively. 3) MBNF based system is much better than the fbank feature based DNN-HMM hybrid system for cross-lingual training, but after semi-supervised training their gaps get close. This indicates that semi-supervised training is more effective on fbank feature based DNN-HMM system.

We note that all systems have used utterance based sequential training [18], using the VLLP supervised data. For the MBNF based DNN system, the recipe is similar to [15], where a 7 hidden layers DNN with 1024 neurons for each layer is trained, using the MBNFs. For the semi-supervised training, the best single system (row 3) is employed to transcribe the unlabeled data.

2) Examplar based acoustic modeling: As shown in Table V, the examplar based acoustic modeling method outperforms the DNN-HMM hybrid method. With only supervised training, the examplar based system has achieved a similar WER as the semi-supervised trained DNN system, i.e. 54.6% (in Table IV). Actually, a little bit trick is used for this kind of system. We

used the logarithmic likelihood scores of the examplar system as inputs to train another one hidden layer neural network. Specifically, the neural network taking 1000 features as input (whose dimension is equal to the number of states of the examplar system), while its soft-max layer is borrowed from the conventional GMM-HMM system.

B. Keyword search performance

1) KWS performance for single systems: Table VI presents the setting and KWS performances of all individual systems, developed using the techniques in Section III and IV, for both *dev* and *eval* keywords lists. In this table, we denote: 1) Monolingual DNN-HMM hybrid as Mono-DNN; 2) Multilingual DNN-HMM hybrid as MDNN; 3) MBNF based DNN-HMM hybrid as MBNF; 4) semi-supervised training as STT. Note that the KWS results, i.e. ATWV, are obtained after applying the well-known KST normalization that is part of the Kaldi recipe [40]. Also note that we did not include the semi-supervised training for MBNF since it only offers little improvement over the supervised MBNF system.

We observe from Table VI, IV, V that ATWV and WER performance are highly correlated. Specifically, we can conclude that: 1) The both multilingual training recipes, i.e. MDNN and MBNF, significantly outperform the monolingual baseline. For example, with the same word-based decoding, the MDNN (S3) and MBNF (S5) systems achieve 9.0% (from 0.2917 to 0.3820) and 12.2% (from 0.2917 to 0.4136) absolute improvement over the baseline Mono-DNN (S1) on *eval* keyword list respectively. 2) MBNF based system (S5 and S6) outperform the corresponding DNN-HMM hybrid system (S3 and S4) 3.2% and 3.3% absolute ATWV on *eval* keyword list respectively. 3) Semi-supervised training is effective for the KWS task. The semi-supervised systems S7 and S8 provide 3.9% and 2.9% absolute improvement over the corresponding supervised training systems, i.e. S3 and S4, on *dev* keyword list. 4) The examplar-based acoustic model generally outperforms the DNN acoustic model for the same setting. Specifically, the examplar-based systems (S9 and S10) outperform the corresponding MBNF system (S5 and S6) from 0.7% to 1.4% absolute ATWV on the two keyword lists.

It can also be seen from Table VI that the subword-based approach is effective, especially for *dev* keyword list. With the same acoustic model training, the subword systems consistently achieve better ATWV than the corresponding word-based systems on the *dev* keyword list. For example, the subword-based MDNN (S4) outperforms the corresponding word-based system S3 by 1.6 % absolute ATWV. On the *eval* keyword list, the subword approach is worse than the word-based approach, but this can be explained by the fact that the OOV rate in *eval* keyword list is lower than in *dev* keyword list (as shown in Table III). In conclusion, the subword approach is shown to be comparable and complementary with the word-based approach.

2) The KWS performance of the classifier-based fusion: In this section, we compare the results of the proposed classifier-based, i.e. the neural network (NN), fusion with the baseline WCombMNZ introduced in [35]. Note that we apply the same KST normalization to the results of the WCombMNZ or the NN-based methods.

TABLE VI
The settings and KWS performances of the baselines and proposed systems

System ID	Training method	Unit	ATWV	
			dev	*eval*
S1	Mono-DNN	Word	0.2517	0.2917
S2	Mono-DNN	Subword	0.2831	0.2878
S3	MDNN	Word	0.3333	0.3820
S4	MDNN	Subword	0.3493	0.3555
S5	MBNF	Word	0.3703	0.4136
S6	MBNF	Subword	0.3845	0.3888
S7	MDNN + SST	Word	0.3725	0.4090
S8	MDNN + SST	Subword	0.3787	0.3920
S9	MBNF + examplar	Word	0.3800	0.4205
S10	MBNF + examplar	Subword	0.3950	0.4028

Since the NN is a supervised method and NIST only provides the reference transcriptions of 10 hours development (dev 10h) data, we decide to split the dev 10h into two sets: the first set is about 7h which is used to train the DNN, and the second set is about 3h to evaluate the fusion. The splitting procedure follows some rules: the speakers in two sets are non-overlapped and the speaker's genders are balanced in both two sets.

We use the searched results of one keyword to train the NN classifier, and then test on other keyword list. For example, in order to test the NN fusion on dev 3h for *eval* keyword list, we use the searched results of KWS systems on dev 7h for *dev* keyword as the training data. The NN has 2 hidden layers and each layer has 200 nodes. Table VII shows the ATWV metric of individual systems (see Table VI for more detail about the setting of each system) as well as two fusion methods on the dev 3h. Note that we did not include the two baseline monolingual systems S1 and S2 into the fusion process since they are much worse than other systems.

TABLE VII
ATWV metric of individual systems as well as two fusion methods WCombMNZ and NN on dev 3h data set

System ID	ATWV	
	dev	*eval*
S3	0.3228	0.3968
S4	0.3185	0.3672
S5	0.3477	0.4003
S6	0.3529	0.3767
S7	0.3507	0.4235
S8	0.3428	0.3772
S9	0.3741	0.4238
S10	0.3726	0.4049
WCombMNZ fusion	0.4399	0.4701
NN fusion	**0.4804**	**0.4971**

It can be seen that both fusion methods significantly outperform the best single system (S9). Specifically, two fusion methods outperform the system S9 by 6.5% and 10.6% absolute ATWV on the dev keyword list. This suggests that our

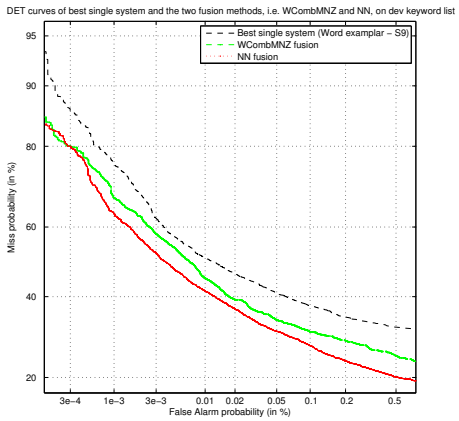

Fig. 3. *The DET curves of the best single system and two fusion methods WCombMNZ and NN on dev keyword list*

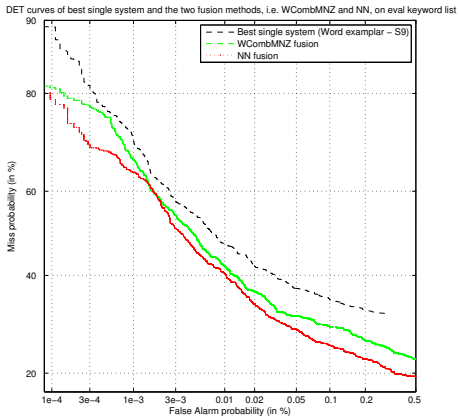

Fig. 4. *The DET curves of the best single system and two fusion methods WCombMNZ and NN on eval keyword list*

subsystem components are complementary.

The second observation is that the NN fusion outperforms significantly the baseline WCombMNZ fusion for both keyword lists: the NN fusion provides 4.1% and 2.7% absolute improvements over the WCombMNZ for *dev* and *eval* keyword lists respectively. Figure 3 and 4 show the DET curves of the two fusion methods as well as the best single system, i.e. the system S9, on *dev* and *eval* keyword lists resepectively. It can be seen that the NN fusion outperforms the WCombMNZ fusion at the whole region for both keyword lists. This can be explained by the fact that the NN helps to incorporate various factors, into the fusion process, hence improves the quality of the fusion scores.

VI. Conclusions

In this work, we investigated various strategies for the very low-resource keyword search task. We first showed that the multilingual and semi-supervised acoustic model training are essential for the very limited resource condition in terms of both the WER and the ATWV result improvement. We also showed that the proposed examplar-based acoustic modeling framework generally outperforms the conventional DNN acoustic modeling framework in the very limited acoustic data condition. Furthermore, we demonstrated that the subword ASR is effective to alleviate the effect of the excessive OOV keyword search issue and its performance is comparable with the word-based couterpart. Finally, the results revealed the proposed classifier-based system fusion consistently outperforms the traditional rule-based fusion approaches.

Acknowledgment

This work is supported by the DSO funded project MAISON DSOCL14045, Singapore. We would like to thank our colleagues in the SINGA team for the valueable disscussions.

References

[1] NIST, "The spoken term detection (std) 2006 evaluation plan," in *http://www.nist.gov/speech/tests/std*, 2006.
[2] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," in *Transaction on Audio Speech and Language Processing*. IEEE, 2011.
[3] S. Parlak and M. Saraclar, "Spoken term detection for turkish broadcast news," in *Proceedings of ICASSP*. IEEE, 2008.
[4] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata application to spoken utterance retrieval," in *Proceedings of HLT-NAACL*. NAACL, 2004.
[5] K. Thambiratmann and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2007.
[6] J. Cernocky et.al., "Search in speech for public security and defense," in *Proceedings of SAFE*. IEEE, 2007.
[7] I. Szoke, J. Cernock, L. Burget, and M. Fapo, "Sub-word modeling of out of vocabulary words in spoken term detection," in *Proceedings of SLT*. IEEE, 2008.
[8] O. Siohan and M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing," in *Proceedings of Interspeech*, 2005.
[9] Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu, "An improved consensus-like method for minimum Bayes risk decoding and lattice combination," in *Proceedings of ICASSP*. IEEE, 2010.
[10] Franktisek Grézl and Martin Karafiát, "Semi-supervised bootstrapping approach for neural network feature extractor training," in *Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2013.
[11] Haihua Xu, Hang Su, Eng-Siong Chng, and Haizhou Li, "Semi-supervised training with bottle-neck feature based dnn-hmm hybrid modeling framework," in *Proceedings of Interspeech*, 2014.
[12] Hang Su and Haihua Xu, "Multi-softmax deep neural network for semi-supervised training," in *Proceedings of Interspeech*, 2015.
[13] NIST, "The open keyword search (std) 2013 evaluation," in *http://www.nist.gov/itl/iad/mig/upload/OpenKWS13-evalplan-v4.pdf*, 2013.
[14] NIST, "The open keyword search (std) 2014 evaluation," in *http://www.nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf*, 2014.
[15] 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 *Proceedings of Interspeech*, 2015.
[16] William Hartmann, Anindya Roy, Lori Lamel, and Jean-Luc Gauvain, "Acoustic unit discovery and pronunciation generation from a grapheme-based lexicon," in *Proceedings of Automatic Speech Recognition and Understanding (ASRU) Workshop*. IEEE, 2013.
[17] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP*, 2013.
[18] Karel Veselý, Arnab Ghoshal, Lukas Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *Proceedings of Interspeech*, 2013.
[19] Frantisek Grézl, Martin Karafiát, and Karel Veselý, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *ICASSP*, 2014.

[20] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Jrmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recogniton," in *Proceedings of ICASSP*, 2014.
[21] T. Deselaers, G. Heigold, , and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in *Proceedings of Interspeech*, 2007.
[22] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. F. Gemmeke, J. R. Bellegarda, , and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," in *Signal Processing Magazine, vol. 29, no. 6*. IEEE, 2012.
[23] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Kernel density-based acoustic model with cross-lingual bottleneck features for resource limited lvcsr," in *Proceedings of Interspeech*, 2014.
[24] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 nist spoken term detection evaluation," in *Proceedings of Interspeech*, 2007.
[25] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proceedings of SIGIR*. IEEE, 2007.
[26] Y. Pan and L. Lee, "Performance analysis for lattice-based speech indexing aproaches using words and subword units," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2010.
[27] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for german spoken term detection," in *Proceedings of ICASSP*. IEEE, 2009.
[28] H. Su, V. T. Pham, Y. He, and J. Hieronymus, "Improvements on transducing syllable lattices to word lattice for keyword search," in *Proceedings of ICASSP*. IEEE, 2015.
[29] R. Wallace, R. Vogt, and S. Sridharan, "Indexing confusion networks for morph-based spoken document retrieval," in *Proceedings of SIGIR,*. IEEE, 2007.
[30] N.F. Chen et.al., "Low-resource keyword search strategies for tamil," in *Proceedings of ICASSP*. IEEE, 2015.
[31] F. Seide and P. Yu, "Vocabulary-independent search in spontaneous speech," in *Proceedings of ICASSP*. IEEE, 2004.
[32] M. Akbacak, D. Vergyri, and A. Stolcke, "open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *Proceedings of ICASSP*. IEEE, 2008.
[33] B. Logan, J. Van Thong, and P. Moreno, "Approaches to reduce the effects of oov queries on indexed spoken audio," in *IEEE Transactions on Audio, Speech, and Language Processing*. IEEE, 2005.
[34] Mathias Creutz and Krista Lagus, "Unsupervised discovery of morphemes," in *In Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, 2002.
[35] J. Mamou et.al., "System combination and score normalization for spoken term detection," in *Proceedings of ICASSP*. IEEE, 2013.
[36] V. T. Pham, N. F. Chen, S. Sivadas, H. Xu, I. F. Chen, C. Ni, E. S. Chng, and H. Li, "System and keyword dependent fusion for spoken term detection," in *Proceedings of SLT*. IEEE, 2014.
[37] D. Karakos et.al., "Score normalization and system combination for improved keyword spotting," in *Proceedings of Automatic Speech Recognition and Understanding (ASRU) Workshop*. IEEE, 2013.
[38] V. Soto, L. Mangu, A. Rosenberg, and J. Hirschberg, "A comparison of multiple methods for rescoring keyword search lists for low resource languages," in *Proceedings of Interspeech*, 2014.
[39] D. Karakos, I. Bulyko, R. Schwartz, S. Tsakalidis, L. Nguyen, and J. Makhoul, "Normalization of phonetic keyword search scores," in *Proceedings of ICASSP*, 2013.
[40] D. Povey et.al, "The kaldi speech recognition toolkit," in *Proceedings of ASRU*, 2011.