DISCRIMINATIVE SCORE NORMALIZATION FOR KEYWORD SEARCH DECISION

Van Tung Pham[1], Haihua Xu[1], Nancy F. Chen[2], Sunil Sivadas[2], Boon Pang Lim[2], Eng Siong Chng[1], Haizhou Li[2].

[1]Nanyang Technological University, Singapore, [2]Institute for Infocomm Research, Singapre
vantung001@e.ntu.edu.sg

ABSTRACT

Many keyword search (KWS) systems make "hit/false alarm (FA)" decisions based on the lattice-based posterior probability, which is incomparable across keywords. Therefore, score normalization is essential for a KWS system. In this paper, we investigate the integration of two novel features, ranking-score and relative-to-max, into a discriminative score normalization method. These features are extracted by considering all competing hypotheses of a putative detection. A metric-based normalization method is also applied as a post-processing step to further optimize the term-weighted value (TWV) evaluation metric. We report empirical improvements over standard baselines using the Vietnamese data from IARPA's Babel program in the NIST OpenKWS13 Evaluation setup.

Index Terms— score normalization, spoken term detection (STD), keyword spotting, confidence estimation, discriminative modeling, under-resourced languages.

1. INTRODUCTION

Recently, more and more data in the form of broadcast news, voice mail, lectures and presentation recordings are being archived. However, those data are often not transcribed, hence cannot be retrieved easily. Hence, keyword search (KWS) [1, 2] is an important area of research. Different from spoken document retrieval [3] – the task to retrieve entire documents of a keyword, KWS aims to find all occurrences of a keyword in a corpus.

Generally, a two-phase approach, namely indexing and search, is utilized for a KWS system. Each audio of the speech corpus is automatically segmented and then passed to a large vocabulary continuous speech recognition (LVCSR) system to produce the corresponding word lattice. These lattices are then indexed using techniques such as weighted finite state transducer (WFST) [4, 5, 6] framework or N-gram indexing [7, 8, 9, 10]. In search phase, keywords in the textual format are searched on the index to produce a list of putative detections.

For each putative detection, KWS systems will make a decision whether it is a hit or a false alarm by comparing the score of the detection to a certain threshold. Currently, many KWS systems [11, 12, 13, 14] use the posterior probability of the keywords computed from the decoded lattices as the scores to make decisions. However, it is observed that the same threshold results in different term weighted value (TWV) performances for different keywords. This is because each detection is affected by various characteristics such as the cost of miss, the cost of false alarm, query length, number of vowels in a specific keyword and context consistence of the keyword. Existing strategies such as keyword specific threshold and sum-to-one [12, 15] have been shown to be effective solutions to this problem. Discriminative score normalization [16, 17, 18, 19, 20, 21] is another approach which aims to normalize scores through discriminative modeling.

In this paper, we investigate the integration of two novel features, namely *ranking-score* and *relative-to-max*, into a discriminative score normalization method. By considering all competing hypotheses of a putative detection, the two novel features are extracted and then combined with other features to train a binary classifier. We also propose to apply a metric-based normalization as a post-processing step to further optimize the TWV evaluation metric.

The paper is organized as follows. In section 2, we discuss our method in relation to prior work. Next, we present discriminative score normalization in section 3. Section 4 introduces the two-stage score normalization in which a metric-based normalization method is applied after discriminative score normalization. Experiments and analysis for the NIST OpenKWS13 Vietnamese data are described in section 5. Finally, section 6 is the conclusion.

2. RELATION TO PRIOR WORK

Score normalization is essential for many research areas such as information retrieval [22] and speaker verification [23]. The work presented here is focused on score normalization for keyword search task. The aim of our work is to estimate normalized scores by taking the characteristics of keywords into account to make the normalized scores comparable across keywords.

One score normalization approach transforms raw posterior probability to a new score to optimize the evaluation metric. By considering the miss and false alarm (FA) cost, researchers proposed keyword specific threshold

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

[12] and sum-to-one [15] normalization techniques to optimize the TWV metric. Keyword specific threshold estimates the specific threshold for each keyword to minimize the decision cost. Sum-to-one boosts score of putative detections of rare keywords based on the TWV characteristic that losing a rare keyword is expensive and missing a frequent keyword is cheap. In this paper, we will call the above two methods as *metric-based* normalization methods.

Another score normalization approach is discriminative score normalization which aims to estimate new scores from features so that the new scores are more discriminative for hit/FA decisions. In this approach, features such as number of vowels, posterior probability and context consistence are used to train a discriminative classifier such as the Multi-Layer Perceptron (MLP) [16, 17], Support Vector Machine [16, 17, 18] or Conditional Random Field [19, 20]. Research shows that the new scores increase correct hit/FA decisions over the raw posterior probability and hence lead to better performance. Although this approach was referred as confidence estimation or discriminative modeling [16, 17, 19], it is another score normalization approach which aims to achieve the implicit normalization through discriminative modeling. Therefore, in this paper, we call this approach as *discriminative score* normalization.

In this paper, we also use MLP as discriminative score normalization method. We further introduce ranking-score and relative-to-max as two novel features for the MLP. We also propose to apply a metric-based normalization as a post-processing step to further optimize the TWV metric.

3. DISCRIMINATIVE SCORE NORMALIZATION

Theoretically, the conventional lattice-based posterior probability represents the posterior probability of a keyword K in the lattice L given the acoustic observation O. A potential drawback of the lattice-based posterior probability lies in the fact that the scores of putative detections of all keywords are treated the same. In general, different keywords exhibit a high diversity in terms of characteristics such as length, number of vowels, etc., implying that the scores of putative detections fall into different ranges.

A more desirable score for a putative detection might be the posterior probability that the detection is correct given the detection, i.e., P(*hit*|*d*), where *hit* represents the event that detection *d* is a hit. The new score should not only depend on the raw posterior probability but also depend on keyword characteristic. This is the motivation to use discriminative score normalization as it allows the integration of any keyword characteristic into a more general framework. In other words, we construct a mapping between set of features to the final score

$$S'_{K,i} = f(f_0, f_1, \ldots), \quad (1)$$

where $S'_{K,i}$ is the desired score of the i^{th} putative detection of keyword K and f_0, f_1,...are features such as raw posterior probability, average query length and number of vowels.

In this paper we choose MLP as a discriminative model. The feature set includes two novel features called *ranking-score* and *relative-to-max* as well as some well-known features: average query length, raw posterior probability, number of vowels. Those features are extracted by considering, for each putative detection, all competing hypotheses of the detection in the corresponding lattice.

Ranking-score: Consider a detection *d* of a single-word keyword which is presented as a tuple (*Lattice L, start-time, duration, posterior probability*), we define *T(d)* as a set of arcs in the lattice *L* that overlap with the mid-point time of *d*. The ranking-score feature is the rank of posterior probability of *d* compared to the posterior probability of all items in *T(d)*. For a keyword *K* that consists of more than one word, $K = K_1 K_2 \ldots K_n$, we infer the ranking of a detection *d* of this keyword using the following formula:

$$\text{ranking-score}(d) = \max_{i=1\ldots n}(\text{ranking-score}(K_i)) \quad (2)$$

It is clear that at the same posterior probability, detections with low *ranking-score* are more likely to be hits. Thus *ranking-score* can provide more information to determine whether a putative detection is a hit or an FA.

Relative-to-max: Again, for a putative detection *d* of a single-word keyword, let *T(d)* be the set of arcs in the lattice *L* that overlap with *d*. The *relative-to-max* of *d* is defined as the relative score of the detection *d* compared to the best score of items in *T(d)*, i.e,

$$\text{Relative-to-max}(d) = \frac{\text{posterior\_probability}(d)}{\max_{h \in T(d)}(\text{posterior\_probability}(h))} \quad (3)$$

For a keyword *K* that consists of more than one word, $K = K_1 K_2 \ldots K_n$, we infer the *relative-to-max* of a detection *d* of this keyword using the following formula:

$$\text{Relative-to-max}(d) = \min_{i=1\ldots n}(\text{relative-to-max}(K_i)) \quad (4)$$

This information indicates how much the best hypothesis in *T(d)* dominate the detection hypothesis *d*. The low *relative-to-max* score means the detection hypothesis is highly dominated by another best hypothesis, which mean it is less likely to be a hit even it can have good *ranking-score*. *Relative-to-max* can provide extra information to *ranking-score* and posterior probability to help make correct hit/FA decisions, hence it is also a good indicator for our discriminative model.

4. TWO-STAGE SCORE NORMALIZATION

The above score normalization methods provide scores from a new perspective that potentially leads to a more informed "hit/FA" decision. We note that, the final target is to optimize the TWV metric. Thus, it is desirable to apply a metric-based score normalization method on the results of the discriminative score normalization procedure to get the final score that optimizes with regards to TWV. With this two-stage scheme, we can take advantages of both discriminative and metric-based score normalization approaches. The two-stage scheme is shown in Figure 1 below. Two state-of-the-art metric-based score normalization methods, keyword specific threshold and sum-to-one, can be used in the second step.

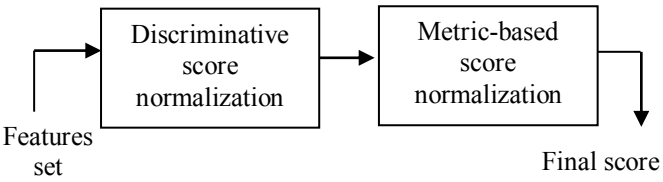

Fig. 1. The two-stage scheme for score normalization

Keyword specific threshold (KST) [12]: For each keyword K, a specific threshold is estimated using the following equation to minimize decision cost:

$$\theta_K = \frac{\beta * N_K}{T+(\beta-1)*N_K} \quad (5)$$

where T is corpus size, β = 999.9 and N_K is number of references of the keyword K that can be approximate by

$$N_K = \sum_{detection} posterior(detection)$$

To allow using global threshold θ for all keywords, an exponential transformation is applied on the raw posterior probability as in the following equation [24]:

$$S'_{K,i} = S_{K,i}^{\frac{\log(\theta)}{\log(\theta_K)}} \quad (6)$$

Sum-to-one normalization (STO)[15]: Sum-to-one method normalizes score to reduce P_{miss} of rare keyword as follows:

$$S'_{K,i} = \frac{S_{K,i}}{\sum_j S_{K,j}} \quad (7)$$

The denominator for rare keyword is small, therefore boosting the normalized score for rare keywords. Thus this normalization helps to reduce P_{miss} for rare keyword.

5. EXPERIMENTS

The keyword search experiments are conducted on the Vietnamese conversational telephone speech used in the NIST OpenKWS13 Evaluation [2]. The training data consists of 80 hours of conversational telephone speech and 20 hours of scripted telephone speech. The development set is 10 hours and the evaluation set is 75 hours. Because NIST only released *evalpart1* (about 15h) of the whole evaluation data, we only conduct evaluation on this part. The evaluation keyword set consists of 4065 keywords and 1309 of those keywords are in the part 1 of the evaluation data.

We used the open-source Kaldi toolkit [25] to build our LVCSR system. We used 13-dim PLP features, and concatenated 9 frames adjacent together to apply LDA, MLLT, and fMLLR transforms, which resulted in 40-dimensions of features to train a deep neural network (DNN) acoustic model. The language model is syllable-based bigram LM with Good Turing Smoothing trained with SRILM toolkit [26]. Details of the system implementation can be found in [27][1]. The final word error rate (WER) of our system is 55.6%.

To evaluate KWS performance, NIST defines the term-weighted value (TWV) [2] which integrates the miss rate and false alarm rate of each query into a single metric and then averages over all queries:

$$TWV(\theta) = 1 - \frac{1}{N}\sum_{term}((P_{miss}(\text{term},\theta) + \beta P_{fa}(\text{term},\theta)) \quad (8)$$

Actual term-weighted value (ATWV) is the TWV of a chosen decision threshold, whereas the maximum term-weighted value (MTWV) is the best TWV found over all the possible values of decision thresholds. The ATWV score is sensitive to the threshold selection thus might lead to uncertainty in comparison between difference experiments. When comparing across ATWV's, it is difficult to know if the difference is caused by different systems or by threshold selection. Therefore, MTWV is used as evaluation metric. In addition to ATWV and MTWV, NIST proposed a detection error tradeoff (DET) curve to evaluate the performance of a KWS system. DET curves are also used for performance evaluation in this paper.

In order to get training data for the MLP model, we augmented the query list of the development set from 200 queries to 460 queries and then searched on those queries to get a list of putative detections. For each putative detection, useful features that are mentioned in section 3 were extracted. The true "hit" or "FA" outputs were obtained by aligning each detection with the reference transcription. This gave us over 40300 examples for the training of the MLP.

5.1. Two novel features for discriminative score normalization

We compare results of the MLP normalization method trained with and without the two new features. We also compare the MLP normalization method with traditional query length normalization proposed in [15, 28] that exponentially transforms score. We denote MLP method without the two new features as MLP-baseline, MLP method with the two new features as MLP-2 and the query length normalization method as QueryLength. The experiment results are shown in Fig. 2 and Table 1. The MLP-baseline method provides only less than 1% absolute improvement compare to the traditional query length normalization. This is not surprising, however, as we only use three features to train the MLP model. When two novel features are integrated into the MLP model, MLP-2 outperforms MLP-baseline and QueryLength by 1.2% and 1.8% absolute respectively.

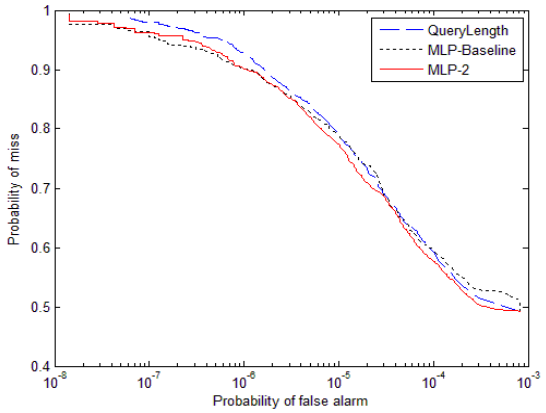

Fig. 2. DET curves for proposed MLP-2 and its baselines

[1] Note that the normalization techniques used in this paper are different from that in [27]

Table 1: MTWV comparison for proposed MLP-2 and its baselines.

Score normalization methods	MTWV
QueryLength	0.3122
MLP-baseline	0.3184
MLP-2 (proposed)	**0.3307**

5.2. Two-stage scheme for score normalization

In this section, we compare the two-stage normalization scheme with the case we only use MLP normalization method. We also compare the results of our proposed method with the standalone keyword specific threshold (denoted as KST) and sum-to-one (denoted as STO) normalization methods that are applied on raw posterior probability. The experiment results are shown in Fig. 3., Fig. 4. and Table 2.

Table 2: MTWV Comparison of proposed MLP-2, keyword specific threshold (KST), sum-to-one (STO), and combinations of score normalization methods.

Score normalization methods	MTWV
MLP-2	0.3307
KST	0.3959
STO	0.3955
MLP-2 + STO	**0.3967**
MLP-2 + KST	**0.4044**

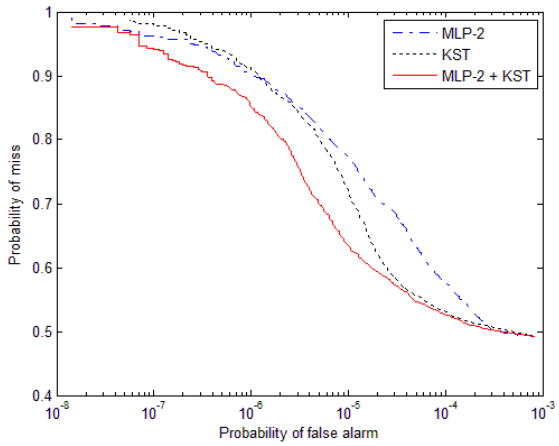

Fig. 3. DET curves of score normalization results using proposed MLP-2, keyword specific threshold (KST) normalization, and their combination (MLP-2 + KST).

From the DET curve in Fig. 3 and Table 2 for keyword specific threshold, it is clear to see the considerable improvement of MLP-2 + KST compare to both baselines, especially MLP-2 alone. Our proposed method outperforms MLP-2 and KST by 7.3 % and 0.85% absolute respectively. For the case of sum-to-one in Fig. 4, our proposed method MLP-2 + STO is at least comparable with STO and still outperforms the baseline MLP-2**.**

The experiments results show that our method is better than the baseline. The reason is that our method takes advantages of both normalization approaches: it considers the characteristics of the keywords as well as the characteristics of TWV evaluation metric.

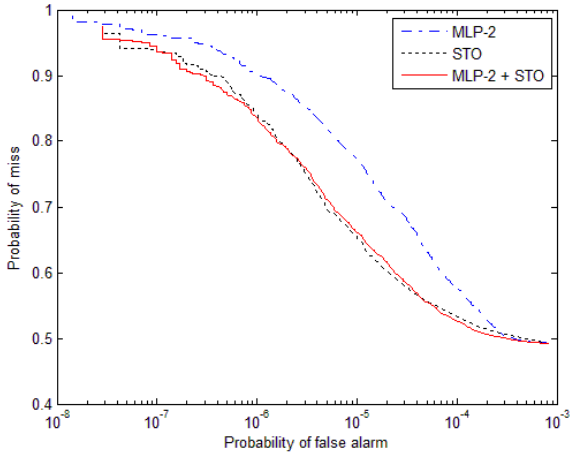

Fig. 4. DET curves of score normalization results using proposed MLP-2, sum-to-one (STO) normalization, and their combination (MLP-2 + STO).

6. CONCLUSION

We have presented a discriminative score normalization method to resolve with the incomparable score problem of the keyword search task. Two novel features, *ranking-score* and *relative-to-max* are integrated into a discriminative classifier to estimate more accurate scores. For each putative detection, the two features are extracted by considering all competing hypotheses of the detection in the corresponding lattice. We show empirically that these features help more precisely estimating final scores. In this paper, we also introduced the two-stage score normalization that uses a metric-based normalization method as a post-processing step. By using this normalization scheme, we can take advantages of both discriminative modeling and metric-oriented optimization normalization approach. For future work, we plan to apply the proposed features and techniques to other languages and conditions of limited language resources.

7. ACKNOWLEDGMENT

We would like to thank Dr. Xiong Xiao of Temasek Labs at Nanyang Technological University for sharing the neural network training code.

8. REFERENCES

[1] NIST, The spoken term detection (STD) 2006 evaluation plan,10th ed., National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, September 2006. [Online]. Available: http://www.nist.gov/speech/tests/std

[2] NIST, Open Keyword Search 2013 Evaluation (OpenKWS13) plan, 4th ed., National Institute of Standards and Technology (NIST), Available:

http://www.nist.gov/itl/iad/mig/openkws13.cfm

[3] T. K. Chia, K. C. Sim, H. Li and H. T. Ng, "Statistical Lattice-Based Spoken Document Retrieval", ACM Transactions on Information Systems, Vol. 28, No. 1, 2010

[4] D. Can, M. Saraclar, "Lattice Indexing for Spoken Term Detection", IEEE Transaction on Audio Speech and Language Processing, Vol. 19, No. 8, 2011

[5] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," in Proc. ICASSP'08, Las Vegas, Nevada, USA, March 2008, pp. 5244–5247

[6] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata application to spoken utterance retrieval," in Proc. HLT-NAACL 2004, Boston, USA, May 2004, pp. 33–40.

[7] K. Thambiratmann and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, no. 1, pp. 346–357, January 2007

[8] J. Cernocky, I. Szoke, I. Fapso, M. Karifiat, M. Burget, L. Kopecky, F. Grezl, P. Schwarz, O. Blembeck, "Search in speech for public security and defense", in proc. SAFE07 pp. 1–7

[9] I. Szoke, J. Cernock, L. Burget, and M. Fapo, "Sub-word modeling of out of vocabulary words in spoken term detection," in proc. SLT, 2008

[10] O. Siohan, M. Bacchiani, "Fast vocabulary-independent audio search using path-based graph indexing", in proc. Interspeech, 2005, pp. 53-56.

[11] I. Szoke, M. Fapso, M. Karafiat, L. Burget, F. Grezl, P. Schwarz, Ondrejlembek, P. Matejka, S. Kontar, and J. Cernocky, "BUT system for NIST STD 2006 - English," in Proc. NIST Spoken Term Detection Evaluation workshop (STD'06). Washington D.C., US: National Institute of Standards and Technology, 2006

[12] D. R. H. Miller, M. Kleber, C. lin Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in Proc. Interspeech'07, Antwerp, Belgium, August 2007, pp. 314–317.

[13] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak,B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in Proc. Interspeech'07, Antwerp, Belgium, 2007, pp. 2393–2396.

[14] J. Mamou, B. Ramabhadran, O. Siohan, "Vocabulary independent spoken term detection," inProc. ACM-SIGIR 2007, Amsterdam, July 2007, pp. 615–622.

[15] J. Mamou, J. Cui, X. Cui, M. J. F. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy and P. C. Woodland, "System combination and score normalization for spoken term detection", in Proc. ICASSP, Vancouver, May 2013, pp. 8273-8276

[16] D. Wang, S. King, J. Frankel and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in Proc. Interspeech'09, Brighton, UK, September 2009, pp. 2139–2142

[17] J. Tejedor, D. T. Toledano, M. Bautista, S. King, D. Wang, and J. Col as, "Augmented set of features for confidence estimation in spoken term detection," in Proc. Interspeech'10, September 2010

[18] H. Lee, Lin-shan Lee, "Improved Spoken Term Detection Using Support Vector Machines Based on Lattice Context Consistence", in Proc. ICASSP 2011

[19] M. S. Seigel, P.C Woodland and M. J. F. Gales "A confidence-based approach for improving keyword hypothesis scores", in Proc. ICASSP 2013.

[20] Carolina Parada, Mark Dredze, Denis Filimonov, and Frederick Jelinek, "Contextual Information Improves OOV detection in Speech," in Proc. NAACL, 2010

[21] J. Tejedor, A. Echeverría, D. Wang, "An evolutionary confidence measurement for spoken term detection", in Proc. CBMI 2011, pp 151-156

[22] M. Fernandez, D. Vallet, P. Castells, "Probabilistic Score Normalization for Rank Aggregation", in proc. 28th European Conference on Information Retrieval 2006, pp. 553-556

[23] V. Hautamäki, T. Kinnunen, F. Sedlak, K.A. Lee, B. Ma and H. Li, "Sparse classifier fusion for speaker verification", IEEE Trans. on Audio, Speech and Language Processing, 21(8), 1622-1631, August 2013

[24] D. Karakos and R.Schwartz, " Score normalization" , Babel PI metting, July 17th 2013

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, " The Kaldi Speech Recognition Toolkit", in Proc. of ASRU 2011

[26] Andreas Stolcke, " SRILM – An extensible language modeling toolkit", in Proc. of Interspeech 2002.

[27] N. F. Chen, S. Sivadas, B. P. Lim, H. G. Ngo, H. Xu, V.T. Pham, B. Ma, H. Li, "Strategies for Vietnamese keyword search", in Proc. of ICASSP 2014.

[28] C. Parada ,A. Sethy, B. Ramabhadran, "Balancing false alarms and hits in Spoken Term Detection", in Proc. ICASSP, Dallas,March, 2010, pp.5286 - 5289