Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 312–320

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia

A novel keyword rescoring method for improved spoken keyword spotting

Ilyes Rebai[a,*], Yassine BenAyed[a], Walid Mahdi[a,b]

[a] *Multimedia InfoRmation system and Advanced Computing Laboratory University of Sfax, Sfax, Tunisia*

[b] *College of Computers and Information Technology, Taif University, Taif, Saudi Arabia*

Abstract

In this paper, we present a spoken KeyWord Spotting (KWS) system which creates a search index from word lattices generated by a deep speech recognizer. Basic KWS systems estimate word posteriors from the lattices and use them to make "correct/false alarm" decisions. The main issue of lattice-based posterior probability is that a putative detection can have very low posterior probability so that the decider fails to detect it and considers it as a false alarm. Therefore, our goal is to enhance the keyword decision by detecting and boosting the score of missed detections. Accordingly, inspired by template matching approach, we propose a new keyword rescoring method. More precisely, detected hits are rescored based on the acoustic similarity and the new score are used then by the decider to make the final decision. Experiments demonstrate that the proposed method potentially leads to more accurate keyword results than the conventional KWS system.

Keywords: keyword spotting, spoken term detection, score normalization, template matching, keyword rescoring

1. Introduction

Speech retrieval is a key information technology which allows for analyzing, indexing and searching for items of data according to a user's information needs. The task of automatically detecting keywords of interest in a stream of continuous speech is known under several names, including "spoken keyword spotting", "audio indexing", and "spoken term detection". Indeed, KeyWord Spotting (KWS) is a key information technology that provides a large-scale access to speech collections. It aims at detecting all occurrences of search keywords of interest within continuous speech. KWS system offers many advantages for data mining tasks such as real-time keyword monitoring and audio document indexing and search [16].

* Corresponding author.
E-mail address: rebai_ilyes@hotmail.fr

10.1016/j.procs.2018.07.265

Early KWS methods, called template-based KWS, were limited by low computing resources and, consequently, research works were limited to the simple tasks such as isolated keyword detection [3,16]. As speech technology matured, more advanced tasks were investigated and more efficient methods were developed, known as acoustic-based KWS [2,5]. However, both methods focused on building custom detectors for pre-defined keywords. In a real-world scenario, a user should be able to perform an open-vocabulary search. Accordingly, during the last years, methods that integrate Automatic Speech Recognition (ASR) and information retrieval has received considerable interest for both predefined and ad-hoc search and has been more successful for audio indexing and search [12,4,9,8,7]. In a typical ASR-based KWS system, the ASR engine would accurately convert speech to text and text retrieval methods would be applied on the transcription output.

In an ideal setting, word lattice is generated for each speech segment in the collection using the ASR engine [11,12]. The posterior probability of each detection's correctness is then estimated directly from the lattices. Thereafter, given a list of search keywords, the search module returns a list of detections and a final decision is then assigned to each detection according to its posterior probability. However, it was observed that making a detection decision on the raw posterior probabilities results in different performances [9,18]. Recently, score normalization strategies such as Keyword Specific Threshold (KST) [12], Sum-To-One (STO) [10] and discriminative techniques [18] (e.g. multilayer Perceptron-based score normalization) have shown that the detection performance could be improved significantly. These strategies aim at computing a new score based on either the raw posterior probabilities or additional features (e.g., keyword length, number of vowels, etc.) so that optimal performance is ensured. Nevertheless, a potential drawback of such a KWS system is that putative detections may have very low probabilities so that the decider fails to detect them even after score normalization. In fact, these detections can be taken from the other alternatives provided by the lattice. They get therefore low posterior probabilities. During the decision stage, true detections with low posterior probabilities are mostly rejected by the KWS system.

In this paper, a KWS system is designed which follows the ASR-based approach and uses word lattices in order to mitigate the ASR deficiencies. We propose a novel score normalization technique that aims to estimate a new score of correctness of each detection. Specifically, instead of using the raw posterior probabilities, a more preferred parameter for an accurate decision is the acoustic features which could be explored for an effective keyword search decision. Motivated by the success of template matching approach which shows high performance in spotting isolated keywords, our idea is to rescore the detections for each keyword based on the acoustic similarity. Experimental results show that ranking the keyword detections based on this new score yields significant improvements in the KWS performance measured by Actual Term Weighted Value (ATWV) and Maximum Term Weighted Value (MTWV) metrics.

The paper is organized as follows: Section 2 aims to define the basic keyword search paradigm. In Section 3, we list score normalization techniques that can be used for improving keyword search. In the next section, we introduce a two-stage search scheme that combines the proposed template matching based keyword rescoring and score normalization technique. Section 5 presents the experimental setup and KWS results. Section 6 ends with conclusions.

2. Keyword Spotting System

As in [12,9], a basic KWS system has mainly four components: an ASR engine, an indexer, a detector, and a decider. During the indexing phase, the ASR system processes a given speech segment and outputs the corresponding transcripts. Specifically, our KWS system uses the word lattice instead of the single-best transcript. Indeed, word lattices instead are used in order to avoid the problem of missing keyword occurrences that are not in the single best transcription. Next, the indexer module takes as input the produced lattices and generated as output an inverted index[1] containing a list of hits for each word in the ASR dictionary. During the search phase, the detector searches for a set of search keywords on the index and generates a scored list of detections. Finally, the decider takes the lists of detected hits and a decision function assigns a threshold and all detections with scores above the threshold are taken as the detection results (YES/NO decision). Figure 1 plots the overall architecture of the KWS system.

[1] An inverted index is a data structure storing a mapping from words to its locations in the search collection.

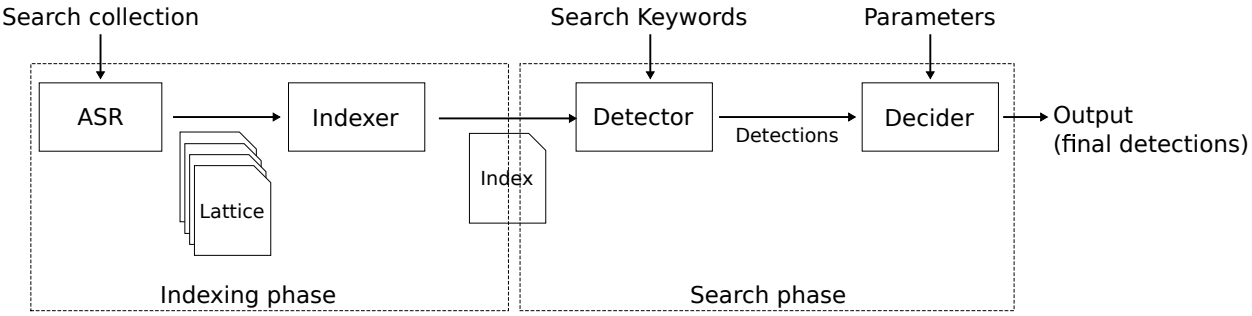


Fig. 1: Keyword spotting system architecture.

KWS accuracy on a given set of search keywords is mainly measured by the so-called Term-Weighted Value (TWV)[1]. This metric is defined by NIST[2], in which the Miss Probability (PMiss), False Alarm Probability (PFA) of each keyword are integrated into a single metric and then averaged over all keywords:

$$TWV(\theta) = 1 - \frac{1}{K}\sum_{w=1}^{K}(PMiss(w,\theta) + \beta\, PFA(w,\theta)) \tag{1}$$

where K is the total number of keywords, T is the total number of trials (e.g., seconds in the audio), and β is a constant, set at 999.9[3]. $PMiss(w,\theta)$ and $PFA(w,\theta)$ are miss and false alarm probabilities for a given keyword w at a specific detection threshold θ.

Two measures are derived from TWV which are the Actual Term-Weighted Value (ATWV) and the Maximum Term-Weighted Value (MTWV). Indeed, ATWV is the TWV of a chosen decision threshold θ, while the MTWV is the best TWV found over all the possible values of decision thresholds. Since ATWV depends on the selected threshold which may lead to uncertain evaluation, MTWV is used to fairly assess the performance of different systems.

3. Score Normalization

Score normalization techniques are used for normalizing the scores across all keywords. Specifically, a decider function is used for computing a decision for each detection so that a detection for keyword w is correct if its posterior score is above a threshold.

3.1. Keyword Specific Threshold

For each keyword w, a specific threshold is estimated using the following formula[12]:

$$\theta^w = \frac{N^w_{true}}{T/\beta + \frac{\beta-1}{\beta}N^w_{true}} \tag{2}$$

where N^w_{true} is an estimate of the expected count of the keyword w in the speech. In real application where the original transcription is absent, N_{true} can be approximated by the expected count for that keyword or is estimated as the sum of the scores for all detections of w. This allows to estimate the threshold in an unsupervised manner.

$$N^w_{true} = \sum_{detection} posterior(detection) \tag{3}$$

[2] https://www.nist.gov/

[3] $\beta = C/V(Pr^{-1} - 1)$ where $C = 0.1$ is the cost of a false detection, $V = 1$ is the value of a correct detection, and $Pr = 10^{-4}$ is the prior probability of a keyword.

To allow using global threshold θ for all keywords, an exponential transformation is applied on the raw posterior probability S:

$$S'_w = S_w^{\left(\frac{log(\theta)}{log(\theta^w)}\right)} \tag{4}$$

Finally, the decider makes a new decision based on the new normalized scores S'.

3.2. Sum-To-One

Sum-to-one technique normalizes score to reduce the miss detection of rare keyword[10,18]. It aims at normalizing the posteriors of each keyword so that the new scores will be in the range of [0..1] and all keywords have then scores that are comparable to each other. In practice, the new score of the i^{th} detection of keyword w is equal to the i^{th} raw posterior probability divided by the sum of the raw posterior probability for all detections of w:

$$S'_{w,i} = \frac{S_{w,i}}{\sum_j S_{w,j}} \tag{5}$$

The denominator is the sum of posteriors for all occurrences. It represents an approximation of the number of occurrences of the keyword. For rare terms, the denominator will be low and therefore the normalized score will be high and can be above the decision threshold[10].

It has been shown that KST function is more effective than STO[17]. Therefore, KST is adopted in this work.

4. Proposed KWS System: Template Matching based Keyword Decision

Our goal is to improve the detection performance and in particular to improve the decision stage which has an intense impact on the final detection result. Score normalizations were explored in the previous works in order to handle the issue of making the lattice-based posterior probability comparable across different keywords[10,9,18]. However, the major issue of such a system is that true detections can have very low probabilities since the search space is expanded by exploring low-quality hypotheses. Therefore, the decider may fail to detect them even after score normalization. Accordingly, our proposed method aims to detect and boost the scores of true detections with low posterior probabilities in order to have an accurate decision.

Our proposed system is mainly based on the basic KWS architecture. Specifically, we used our proposed ASR system[14] in order to generate the word lattices. Our system consists on combining multiple deep acoustic models instead of using a single acoustic model as used in the standard approaches. It has the advantage of exploring weak and strong models in a common framework, in which the resulting of the ensemble is generally more accurate than any of the individual models that compose the ensemble. Further details of our ASR system can be found in[14]. Next, the Weighted Finite State Transducer (WFST) based lattice indexation algorithm, proposed by Can and Saraclar[4], is used as the core of the indexer module. Then, a basic detector module is applied which produces as output all occurrences of each input keyword spotted in the index. Each detection is presented as a tuple (utterance ID, start-time, end-time, posterior probability). Finally, our proposed two-stage decision component is used to perform the final decision. It aims at combining a novel keyword rescoring module with the KST based score normalization technique in order to get a more accurate decision. Figure 2 shows the process of the proposed method.

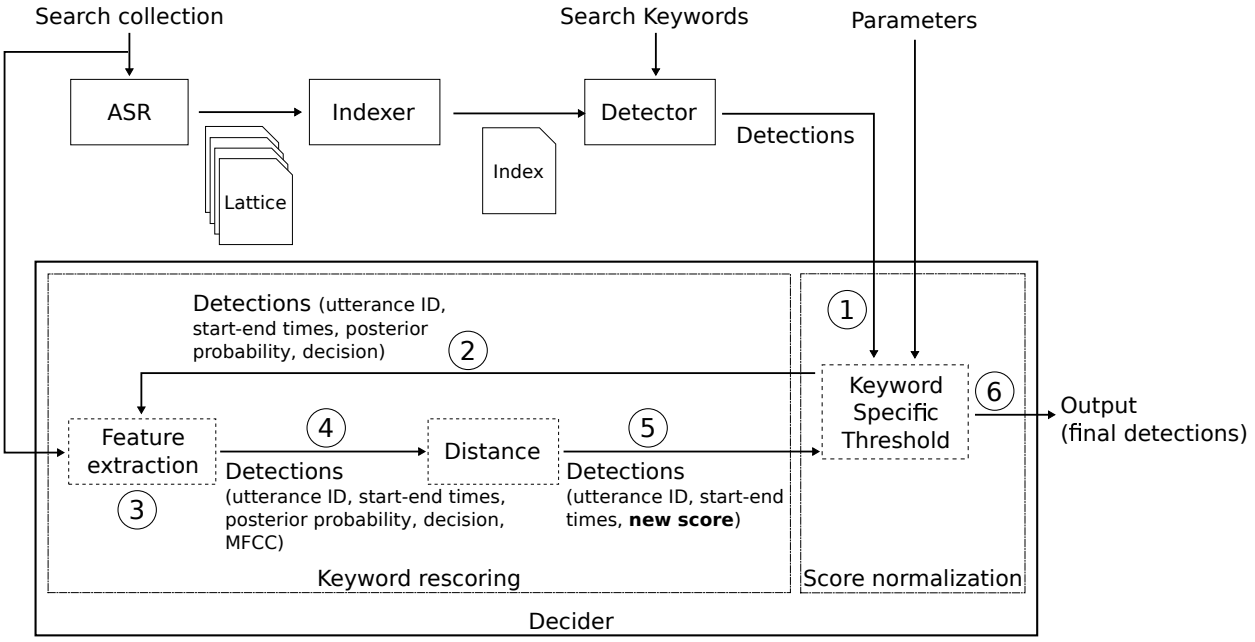


Fig. 2: Keyword spotting based on keyword rescoring method.

The main motivation behind our proposed decision component is the success of template matching approach [3,19,6] in detecting isolated spoken keywords and the effectiveness of score normalization in calibrating the scores of all keywords. In fact, a template matching approach is used to find a match between the template and the object (i.e., keyword) to be detected in the parametric space (i.e., acoustic features). Specifically, it consists of measuring the similarity between multiple samples of a specific keyword, referred to as templates, and a test segment. In this regards, we handle the task of keyword decision as isolated keyword spotting and decision. Accordingly, we propose to replace the raw posterior probabilities by the scores computed using the template matching approach. To perform that, for each keyword, we consider the accepted detections as templates and the rejected detections as detection objects and we compute a simple similarity score between the templates and the objects. The similarity is calculated using a distance metric. The new scores are then normalized across all keywords using KST and the decision threshold is finally applied to assign the final detection result. In the following, we detail the process of the proposed two-stage decision component:

1. The list of detections goes through the decider to get the initial decision.
2. For each detection, utterance ID, start-time, end-time, posterior probability together with a target variable denoting whether the detection is a “true positive” or a “false alarm” are given as input to the feature extraction component.
3. The start-time and end-time information are used to extract the corresponding acoustic features (MFCC) from the input speech.
4. The tuples (utterance ID, start-time, end-time, posterior probability, decision, MFCC) are given as input to the distance component.
5. For each keyword, detections with YES decision are used as templates, and the acoustic similarity score between the templates and each rejected detection (i.e., detection with NO decision) is used as a new score.
6. The new list of detections proceeds through the decision component to make a new decision. The decider recompute the global threshold based on the new scores and all detections with scores above the threshold are taken as the final result.

The distance stage consists on computing the similarity between the detection templates and the rejected detections. The similarity is calculated using a distance metric. It has been shown in template matching based works that increasing the number of templates usually leads to performance improvements[6,19]. Accordingly, a different number of templates are evaluated in this work: 3, 5, 7, and all true detections (i.e., YES detections). Besides that, three different distance metrics are explored in this work which are Cosine, Euclidean, and Standardized-Euclidean:

$$D_{Cosine}(f_1, f_2) = 1 - \frac{f_1 \cdot f_2}{\|f_1\| \|f_2\|} \tag{6}$$

$$D_{Euclidean}(f_1, f_2) = (f_1 - f_2)(f_1 - f_2) \tag{7}$$

$$D_{S\,Euclidean}(f_1, f_2) = (f_1 - f_2)V^{-1}(f_1 - f_2) \tag{8}$$

where f_1 and f_2 correspond to two MFCC features. V is a diagonal matrix whose j^{th} diagonal element is $S(j)^2$, where S is the vector of standard deviations.

5. Experiments

5.1. Setup

The keyword search experiments are conducted on French. The speech corpus consisted of read speech files collected from audiobooks that are part of LibriVox projet[4]. We divided the data by speakers to make sure that speakers appear in one set will not appear in the other sets. The training data consisted of 46 hours, while the test data on which we report performance is of the order to 3 hours. The evaluation keyword set consists of around 40 keywords and the total number of occurrences is about 1000. We have selected the most frequent keywords from the test data. Each keyword is either a single word or composed of two words.

For training the ASR engine, we used 13 normalized MFCC features and augmented with time derivatives, i.e. the first and second derivatives. The obtained 39-MFCC features were concatenated with one left and one right adjacent frames, along with 100 i-vector features, which resulted in 217 feature vector used to train a Deep Neural Network (DNN) acoustic models. Multiple DNN models were trained on different deformed speech data and fused into a single framework using a simple averaging strategy[5]. A tri-gram language model was trained on the training transcripts along with additional data collected from different source using SRILM toolkit[6 15]. The final Word Error Rate (WER) of our system is 20.89%.

We build our KWS pipeline using the open source Kaldi toolkit [7 13]. This toolkit includes all the necessary tools for building and using a basic KWS system. It provides the lattice indexation algorithm of[4] which is the most efficient indexation technique.

To evaluate KWS performance, ATWV/MTWV which integrates the miss rate and false alarm rate as the main evaluation metrics. In addition, a Detection Error Tradeoff (DET) curve is also used to evaluate the performance of a KWS system. We used the scoring tool "Framework for Detection Evaluation" (F4DE) toolkit[8] to compute ATWV, MTWV and the DET curve.

[4] LibriVox is a group of volunteers who read and record public domain texts creating of approximately 25000 public domain audiobooks for download. Link: https://librivox.org

[5] The output of the ensemble is obtained as the average posterior probabilities produced by N acoustic models.

[6] http://www.speech.sri.com/projects/srilm/

[7] https://github.com/kaldi-asr/kaldi

[8] https://github.com/usnistgov/F4DE

5.2. Results

The aim of the present experiments is to evaluate the performance of the proposed KWS system. The first set of experiments leads to the definition of the optimal parameters of the proposed keyword rescoring method, namely the number of templates and the distance metric. The second set of experiments are conducted to compare the performance of our system over the basic one.

5.3. Keyword Rescoring Results

Table 1 presents the keyword detection measured by ATWV and MTWV when different numbers of templates (i.e., detections with YES decision) are given.

Table 1: Comparison of different number of templates; denoted as Tpls.

	3-Tpls	5-Tpls	7-Tpls	All-Tpls
ATWV	0.5456	0.5485	**0.5499**	0.5496
MTWV	0.6415	0.6443	**0.6458**	0.6455

It could be seen that increasing the number of templates to the distance comparison leads to performance improvements. For instance, increasing the number of templates from three to seven samples improves the detection. This observation is consistent with template matching based studies which demonstrated that increasing the number of templates usually leads to performance improvements[19]. This could be explained by the fact that the detection is highly affected by various variabilities (e.g., speaker variability, recording conditions, etc.) between the template and the test segment. To gain some robustness, a simple and effective solution is to increase the number of templates which increases subsequently the system reliability with respect to changes in the recording conditions.

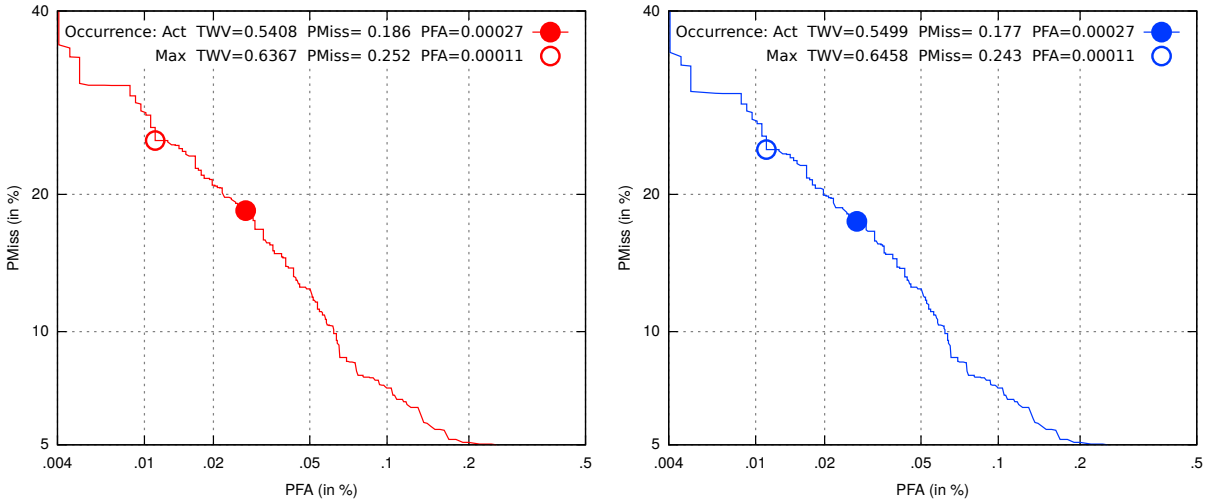
Nonetheless, using all true detections as templates degrades the performance. This could be explained by the fact that additional templates may include false alarms which may negatively impact on the comparison between the templates and the test keywords.

Next, we compared the predefined distance metrics: cosine, Euclidean, and standardized Euclidean. In this experiments, seven templates are used as templates as it gave the best performance. Table 2 presents the obtained results. It can be seen that keyword re-scoring based on cosine metric achieves the best performance over Euclidean and standardized Euclidean metrics with ATWV and MTWV of 0.5499 and 0.6458 respectively. For the remaining experiments, we fix the cosine distance metric and seven templates.

Table 2: Comparison of different distance metrics.

	Cosine	Euclidean	SEuclidean
ATWV	**0.5499**	0.5464	0.5441
MTWV	**0.6458**	0.6420	0.6402

The performance superiority of Cosine over Euclidean and SEuclidean metrics is due to the fact that Cosine similarity is generally used as a metric for measuring the distance between two vectors and not the magnitude: two vectors with the same direction have a Cosine similarity of 1, independent of their magnitude. However, Euclidean and its variant SEuclidean measure the magnitude (L2-norm) of two vectors instead of direction. In our case, the distance between the templates and the keyword segments is more important than magnitude which explains the advantage of Cosine distance over Euclidean and SEuclidean metrics.



(a) Baseline KWS (b) Proposed KWS

Fig. 3: KWS systems curves.

5.4. Evaluation Results

To compare the proposed KWS based on keyword re-scoring method and the basic KWS based on only score normalization, the following performance measures were calculated: PMiss, PFA, ATWV, and MTWV. Figure 3 plots the DET curves for both KWS systems. KWS system based on the score normalization is referred to as the *Baseline KWS*.

Figure 3 reveals the following observations. First, the positive influence of the re-scoring method is demonstrated for all types of performance measures: PMiss, PFA, ATWV, and MTWV. All performance measures for the KWS system based on the proposed keyword re-scoring method are significantly better than the results of the baseline system. For instance, the proposed method provides a 1% absolute improvement in terms of ATWV and MTWV compared to the conventional score normalization technique.

Second, by analyzing the PMiss and PFA measures, the proposed KWS system improves the detection of correct keywords over the baseline yielding a PMiss of 0.243, while both KWS systems achieve the same PFA. This observation demonstrates that missed detections are perfectly detected and re-scored by the proposed method.

Finally, these results prove the effectiveness of the proposed method in improving the keyword search performance. By replacing the raw posterior probabilities with acoustic similarity scores, the decider becomes more accurate in detecting the keywords.

6. Conclusion

In this paper, we presented a spoken keyword spotting system. This system is based on the speech recognition technology and text-matching techniques to spot keywords of interest. Our study mainly aims to address the challenge of improving the keyword detection performance and particularly the keyword decision. The basic decider is primarily based on posterior probabilities generated by the lattice, which may fail to detect accurately putative detections with low posterior probabilities. Accordingly, our objective is to enhance the decision by boosting the score of missed detections.

We introduced a two-stage decision scheme which couple a new keyword rescoring module and score normalization technique. The proposed keyword rescoring method is motivated by the success of QbyE keyword spotting. The idea is to compute a new decision score based on the acoustic similarity of the detected keywords. Experimental results reveal the good performance and advantage of the proposed method. Experiments showed that the keyword rescoring detects properly the missed detections, which is the main objective of our proposal. This demonstrates that

the proposed method is efficient in reducing miss detection rate and a significant performance improvement is thus achieved.

References

1. Open keyword search 2013 evaluation (openkws13) plan, national institute of standards and technology (nist). https://www.nist.gov/sites/default/files/documents/itl/ iad/mig/OpenKWS13-EvalPlan.pdf. 2013.
2. Yassine Benayed, Dominique Fohr, Jean Paul Haton, and Gérard Chollet. Confidence measures for keyword spotting using support vector machines. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–I. IEEE, 2003.
3. John S Bridle. An efficient elastic-template method for detecting given words in running speech. In *Brit. Acoust. Soc. Meeting*, pages 1–4, 1973.
4. Doğan Can and Murat Saraclar. Lattice indexing for spoken term detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(8):2338–2347, 2011.
5. Guoguo Chen, Carolina Parada, and Georg Heigold. Small-footprint keyword spotting using deep neural networks. In *International Conference on, Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4087–4091, 2014.
6. Guoguo Chen, Carolina Parada, and Tara N Sainath. Query-by-example keyword spotting using long short-term memory networks. In *International Conference on, Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5236–5240, 2015.
7. Nancy F Chen, Boon Pang Lim, Chongjia Ni, Haihua Xu, Mark HasegawaJohnson, Wenda Chen, Xiong Xiao, Sunil Sivadas, Eng Siong Chng, Bin Ma, et al. Low-resource spoken keyword search strategies in georgian inspired by distinctive feature theory. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 1322–1327. IEEE, 2017.
8. Nancy F Chen, Chongjia Ni, I-Fan Chen, Sunil Sivadas, Haihua Xu, Xiong Xiao, Tze Siong Lau, Su Jun Leow, Boon Pang Lim, Cheung-Chi Leung, et al. Low-resource keyword search strategies for tamil. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5366–5370. IEEE, 2015.
9. Damianos Karakos, Richard Schwartz, Stavros Tsakalidis, Le Zhang, Shivesh Ranjan, Tim Tim Ng, Roger Hsiao, Guruprasad Saikumar, Ivan Bulyko, Long Nguyen, et al. Score normalization and system combination for improved keyword spotting. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 210–215. IEEE, 2013.
10. Jonathan Mamou, Jia Cui, Xiaodong Cui, Mark JF Gales, Brian Kingsbury, Kate Knill, Lidia Mangu, David Nolden, Michael Picheny, Bhuvana Ramabhadran, et al. System combination and score normalization for spoken term detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8272–8276. IEEE, 2013.
11. Jonathan Mamou, Bhuvana Ramabhadran, and Olivier Siohan. Vocabulary independent spoken term detection. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 615–622, 2007.
12. David RH Miller, Michael Kleber, Chia-Lin Kao, Owen Kimball, Thomas Colthurst, Stephen A Lowe, Richard M Schwartz, and Herbert Gish. Rapid and accurate spoken term detection. In *Eighth Annual Conference of the International Speech Communication Association*, 2007.
13. Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number EPFL-CONF-192584. IEEE Signal Processing Society, 2011.
14. Ilyes Rebai, Yessine BenAyed, Walid Mahdi, and Jean-Pierre Lorré. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 112:316–322, 2017.
15. Andreas Stolcke. Srilm-an extensible language modeling toolkit. In *Seventh international conference on spoken language processing*, 2002.
16. Albert JK Thambiratnam. *Acoustic keyword spotting in speech with applications to data mining*. PhD thesis, Queensland University of Technology, 2005.
17. Yun Wang and Florian Metze. An in-depth comparison of keyword specific thresholding and sum-to-one score normalization. Technical report, Carnegie Mellon University, 2014.
18. Haihua Xu, Nancy F Chen, Sunil Sivadas, Boon Pang Lim, Eng Siong Chng, Haizhou Li, et al. Discriminative score normalization for keyword search decision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7078–7082. IEEE, 2014.
19. Yaodong Zhang and James R Glass. Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams. In *IEEE Workshop on Automatic Speech Recognition & Understanding, 2009. ASRU 2009*, pages 398–403. IEEE, 2009.