IMPROVED SPOKEN TERM DETECTION WITH GRAPH-BASED RE-RANKING IN FEATURE SPACE

Yun-Nung Chen [†], *Chia-Ping Chen* [#], *Hung-Yi Lee* [#], *Chun-An Chan* [#], *and Lin-Shan Lee* [†#]

[†] Graduate Institute of Computer Science and Information Engineering
[#] Graduate Institute of Communication Engineering
[†#] National Taiwan University, Taiwan

vivian.ynchen@gmail.com, coward7652@yahoo.com.tw

ABSTRACT

This paper presents a graph-based approach for spoken term detection. Each first-pass retrieved utterance is a node on a graph and the edge between two nodes is weighted by the similarity between the two utterances evaluated in feature space. The score of each node is then modified by the contributions from its neighbors by random walk or its modified version, because utterances similar to more utterances with higher scores should be given higher relevance scores. In this way the global similarity structure of all first-pass retrieved utterances can be jointly considered. Experimental results show that this new approach offers significantly better performance than the previously proposed pseudo-relevance feedback approach, which considers primarily the local similarity relationship between first-pass retrieved utterances, and these two different approaches can be cascaded to provide even better results.

Index Terms— spoken term detection, re-ranking, pseudo-relevance feedback (PRF)

1. INTRODUCTION

Spoken term detection is to return a list of spoken utterances containing the term requested by the user. Conventional spoken term detection usually includes two stages: the speech recognition system first transcribes spoken utterances into lattices, and then the search engine looks through all lattices for possible presence of the query term [1, 2]. However, in this process much of the information in the acoustic signals may be lost in the stage of speech recognition, especially when the acoustic and language models used are not well matched to the speech signals in the archive to be retrieved, which naturally results in degraded recognition accuracy and poor detection performance. Although many efficient approaches [3, 4] have been proposed to enhance the detection performance due to the relatively poor recognition output, proper use of the feature space information which may be lost during recognition is definitely useful.

Pseudo-relevance feedback (PRF) has been previously borrowed from text information retrieval and successfully applied to the spoken term detection [5, 6]. In this approach, after the first-pass retrieval, a pseudo-relevant utterance set is selected from the first-pass returned list and assumed to be relevant, and the similarity between each first-pass retrieved utterance and this pseudo-relevant utterance set is computed in the feature space and integrated with the original relevance scores for re-ranking the first-pass retrieved utterances. This paper moves one step forward with graph-based re-ranking in feature space. The basic idea is that utterances similar to more utterances with higher relevance scores should be given higher scores, and this concept can be realized by re-ranking over a graph. In this way the global structural information of the first-pass retrieved utterances can be better considered. This approach is similar to the very successful PageRank [7] used to rank the text pages, which considers the relation between every two pages and computes a converged relevance score for each page. Similar concept has been formed useful in video search, in which the similarity between each pair of videos was used to formulate the ranking problem over a graph [8].

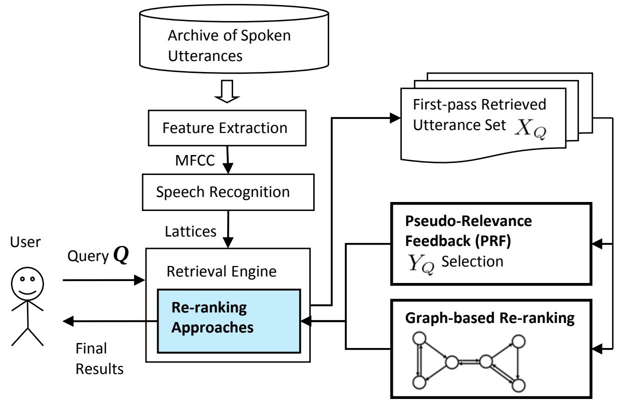

Fig. 1. *The complete framework for the proposed approach*

2. PROPOSED FRAMEWORK

The proposed framework is shown on the Fig. 1. The left half of Fig. 1 is the conventional spoken term detection. We extract MFCC from the spoken utterances in the archive, and translate them into lattices by speech recognition. When the user enters a query Q, the retrieval engine searches over all lattices to find those utterances containing the query Q as the first-pass returned list X_Q ranked by the relevance score $S_Q(x)$. The relevance score $S_Q(x)$ of an utterance x with respect to the query Q is defined as

$$S_Q(x) = \sum_{word(a)=Q} P(a|x), \tag{1}$$

where a is any arc in the lattice of x, $word(a)$ is the word hypothesis of a and $P(a|x)$ is the posterior probability. The first-pass returned list is not shown to user at this stage.

The right half part of Fig. 1 is the proposed approach. We evaluate the similarity between each pair of first-pass retrieved utterances and use it to construct a graph for the first-pass retrieved

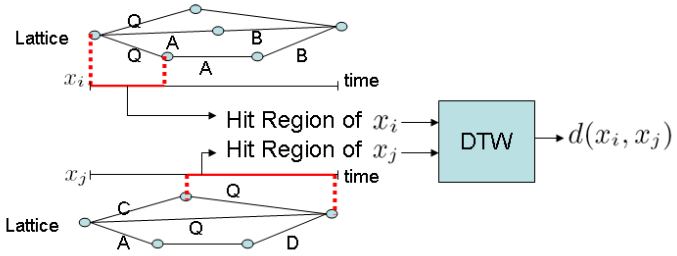

Fig. 2. *The definition of "hit region" (the red part) of an utterance x_i and the distance $d(x_i, x_j)$ between two utterances x_i and x_j. The hit region of an utterance x_i is the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query Q with the highest posterior probability in the lattice.*

utterances, on which re-ranking is performed . First we define the "hit region", the most possible occurrence of query Q in the utterance, as the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query term Q with the highest posterior probability in the lattice. The basic idea here is that if an utterance has a "hit region" very similar to those of utterances with higher relevance scores, it is more likely to be relevant, so its relevance score should be increased. Therefore, we define the distance $d(x_i, x_j)$ between two utterances x_i and x_j given a query Q between the "hit regions" of x_i and x_j, as shown on Fig. 2 [9]. The similarity $sim(x_i, x_j)$ between x_i and x_j is then defined accordingly in Section 3. With the similarity between each pair of utterances in the first-pass returned list X_Q, we then construct a graph for the utterances and apply the graph-based algorithm on the graph to evaluate the new relevance scores for the utterances by considering the relevance scores of similar utterances. Note that different from the previous pseudo-relevance feedback method, the global structure of similarity between the utterances is better considered with the help of the graph in this approach.

3. GRAPH-BASED RE-RANKING

Unlike the pseudo-relevance feedback, the proposed method doesn't need the pseudo-relevant utterance set, because it uses the global structure of all first-pass retrieved utterances. We formulate the re-ranking problem on a directed graph, in which each first-pass retrieved utterance is a node and the edges between them are weighted by the similarity evaluated in the feature space. We initially define two directed edges between each pair of nodes with two directions, both weighted by the similarity between them. We then delete some directed edges by keeping only the top K outgoing edges with the highest weights for each node. A simplified example for such a graph is in Fig. 3. In the above, the similarity between the utterances x_i and x_j is defined as

$$sim(x_i, x_j) = 1 - \frac{d(x_i, x_j) - d_{min}}{d_{max} - d_{min}}, \tag{2}$$

where $d(x_i, x_j)$ is obtained with DTW [9] as in Fig. 2 mentioned above, and d_{max} and d_{min} are the largest and smallest values of $d(x_i, x_j)$ for all pairs of first-pass retrieved utterances for the query Q. We normalize this similarity for an utterance x_i (node i) by the total similarity for x_i and all its neighbors connected by outgoing edges from x_i to produce the weight $p(i, j)$,

$$p(i, j) = \frac{sim(x_i, x_j)}{\sum_{x_k \in A_i} sim(x_i, x_k)}, \tag{3}$$

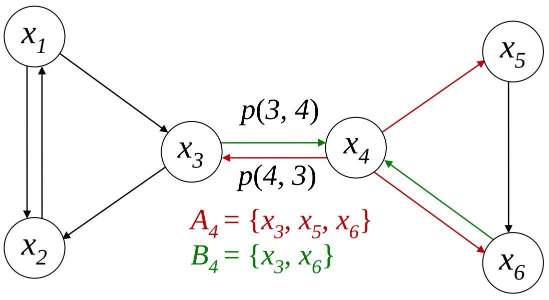

Fig. 3. *A simplified example of the graph considered. Each first-pass retrieved utterance is represented as a node on the graph. A_i and B_i are the neighbors of the node x_i connected respectively by outgoing and incoming edges.*

where A_i is the set of the top K neighbors connected to x_i by the top K outgoing edges of x_i. With this directed graph constructed, we then consider the global structure of the graph to compute the new relevance scores for each utterance (node) by properly integrating the scores of its neighbors (similar utterances). This can be done with at least two different approaches below.

3.1. Random Walk

This approach has been applied to video re-ranking [8]. $v_1(i)$ is the new score defined for node i, which is the interpolation of two scores, the normalized relevance score $r(i)$ for node i and the score contributed by all neighbors j of node i weighted by $p(j, i)$ as defined in (3),

$$v_1(i) = (1 - \alpha)r(i) + \alpha \sum_{x_j \in B_i} p(j, i)v_1(j), \tag{4}$$

where α is the interpolation weight, B_i is the set of neighbors connected to node i via incoming edges, and

$$r(i) = \frac{S_Q(x_i)}{\sum_{x_j \in X_Q} S_Q(x_j)} \tag{5}$$

is the normalized relevance score of utterance x_i, $S_Q(x)$ is as defined in (1), and X_Q is the first-pass retrieved utterance set. Equation (4) can be solved with the approach very similar to that for the PageRank problem [7]. Let $\mathbf{v_1} = [v_1(i), i = 1, 2, ..., L]^{\mathbf{T}}$ and $\mathbf{r} = [r(i), i = 1, 2, ..., L]^{\mathbf{T}}$ be the column vectors for $v_1(i)$ and $r(i)$ for all utterances x_i in the first-pass retrieved set X_Q, where L is the total number of utterances in X_Q and $\mathbf{T}$ represents transpose. Equation (4) then has a vector from below,

$$\begin{aligned} \mathbf{v_1} &= (1 - \alpha)\mathbf{r} + \alpha \mathbf{P} \mathbf{v_1} \\ &= ((1 - \alpha)\mathbf{r}\mathbf{e}^{\mathbf{T}} + \alpha \mathbf{P})\mathbf{v_1} = \mathbf{P_1}\mathbf{v_1}, \end{aligned} \tag{6}$$

where $\mathbf{P}$ is an $L \times L$ matrix of $p(j, i)$, and $\mathbf{e} = [1, 1, ..., 1]^{\mathbf{T}}$ is an L-dimension vector with all components being 1. Because $\sum_i v_1(i) = 1$ from (4) and (5), $\mathbf{e}^{\mathbf{T}}\mathbf{v_1} = 1$.

It has been shown that the solution $\mathbf{v_1}$ of (6) is the dominant eigenvector of $\mathbf{P_1}$ [10], or the eigenvector corresponding to the largest absolute eigenvalue (which is 1) of $\mathbf{P_1}$. The solution $v(i)$ can then be integrated with the original relevance score $S_Q(x)$ for re-ranking,

$$\hat{S_Q}(x_i) = S_Q(x_i)(v_1(i))^{\delta}, \tag{7}$$

where δ is a weighting parameter.

Table 1. *The MAP results for the first-pass baseline, pseudo-relevance feedback (PRF), the proposed graph-based re-ranking and cascade approach, respectively for three sets of acoustic models. ($N = 7, M = 15$ for PRF1, PRF2, and $\alpha = 0.9$ for G1, G2)*

	Methods	**SI**		**MLLR**		**SD**	
		MAP	Impr.	MAP	Impr.	MAP	Impr.
(a)	First-Pass	45.47	-	55.54	-	73.52	-
(b)	Pseudo-Relevance Feedback with Direct Selection (PRF1)	52.10	6.63	61.59	6.05	75.78	2.26
(c)	Pseudo-Relevance Feedback with Min-distance Selection (PRF2)	52.63	7.16	64.07	8.53	76.30	2.78
(d)	Graph-based with Random Walk (G1)	53.42	7.95	63.78	8.24	76.71	3.19
(e)	Graph-based with Modified Random Walk (G2)	54.37	8.90	66.82	11.28	**78.44**	**4.92**
(f)	Cascade 1: PRF2 + G1	53.39	7.92	64.36	8.82	76.27	2.75
(g)	Cascade 2 : PRF2 + G2	**57.75**	**12.28**	**67.38**	**11.84**	77.47	3.95
(h)	Max Relative Improvement (%)	+27.01		+22.04		+6.69	

3.2. Modified Random Walk

This approach is very similar to the Random Walk approach presented above, except $p(j, i)$ in (4) is replaced by $p(i, j)$ and the set B_i for all neighbors connected by incoming edges is replaced by A_i for neighbors connected by outgoing edges. Because $\sum_i v_2(i) = 1$ is not necessarily true in general, we add the normalizing factor λ, and have

$$v_2(i) = \frac{1}{\lambda}((1-\alpha)r(i) + \alpha \sum_{j \in A_i} p(i,j)v_2(j)). \quad (8)$$

Equation (8) can be similarly represented as above,

$$\begin{aligned} \mathbf{v_2} &= \frac{1}{\lambda}((1-\alpha)\mathbf{r} + \alpha\mathbf{P^T}\mathbf{v_2}) \\ &= \frac{1}{\lambda}((1-\alpha)\mathbf{r}\mathbf{e^T} + \alpha\mathbf{P^T})\mathbf{v_2} = \mathbf{P_2}\mathbf{v_2}. \end{aligned} \quad (9)$$

According to Perron-Frobenius Theory, it can be shown that adding a normalized factor λ here leads to the unique solution of (9), the dominant eigenvector of $\mathbf{P_2}$, very similar to Random Walk. We then similarly integrate the scores $v_2(i)$ with the original relevance scores as in (1) and re-rank the utterances.

4. EXPERIMENTS

4.1. Experimental Setup

We used a corpus of 33 hours of recorded lectures for a course offered in National Taiwan University produced by a single instructor primarily in Mandarin Chinese as the testing archive to be retrieved, which is quite noisy and spontaneous. A lexicon of 10.7K words and a tri-gram language model trained with 600M of news data were used in speech recognition. Mean average precision (MAP) was used as the measure for retrieval performance evaluation. 162 Chinese queries were manually selected in the tests, each being a single word.

In order to test the performance of the proposed approach with respect to acoustic models of different matched conditions, we used three sets of acoustic models:

1. The Speaker Independent model (SI) trained by 24.6 hours of read speech produced by 100 male and 100 female speakers.
2. The MLLR model (MLLR) adapted from the above SI model with 500 utterances taken from the training set of the lecture corpus used here.
3. The Speaker Dependent model (SD) trained with 12 hours of the training set of the lecture corpus used here, all produced by the same speaker as those to be retrieved.

In all the three sets of acoustic models, we trained 4602 state-tied triphone models. Each triphone model had 5 states, each with 24 Gaussian mixtures. The recognition accuracy was 50.26%, 62.55% and 81.34% respectively for the SI, MLLR and SD models described above.

4.2. Evaluation Results

The results for the first-pass retrieval for the three sets of acoustic models are listed in row (a) of Table 1 as the first baseline. Clearly the performance is heavily dependent on the quality of the acoustic models.

4.2.1. Pseudo-Relevance Feedback (PRF)

The pseudo-relevance feedback (PRF) approach proposed earlier [5] was used as the second set of baselines. In this approach, a pseudo-relevant utterance set Y_Q was selected out of the first-pass retrieval results X_Q for a query Q, and the similarity between each utterance in the first-pass returned list and this set was computed and integrated with the original relevance score. Two versions were tested here.

- Direct Selection: It simply used the top N utterances in X_Q as Y_Q, and the results are listed in row (b) of Table 1 (PRF1).
- Minimum-distance Selection: It was more complicated. Top M utterances ($M > N$) in X_Q were first picked up to form a set, and the distance between each utterance in X_Q and this set was evaluated. The N utterances with minimum distance obtained in this way was Y_Q. The results are listed in row (c) of Table 1 (PRF2).

In both cases the similarity between an utterance x_i and the set Y_Q is evaluatd by

$$D(x_i, Y_Q) = \sum_{x_j \in Y_Q} d(x_i, x_j)^2, \quad (10)$$

$$SIM(x_i, Y_Q) = 1 - \frac{D(x_i, Y_Q) - D_{min}}{D_{max} - D_{min}}, \quad (11)$$

where $D(x_i, Y_Q)$ is the total distance between x_i and all utterances in Y_Q, and (11) is very similar to (2).

We see that both approaches in rows (b)(c) are much better than the first-pass results in row (a) regardless of the quality of the acoustic models. Also, the second approach of Pseudo-Relevance Feedback with Minimum-distance Selection (PRF2) in row (c) performed always better than the first approach of Pseudo-Relevance Feedback with Direct Selection (PRF1), obviously because the pseudo-relevant utterance set is more reliable for PRF2. Rows (b)(c) serve as the next two baselines to be compared.

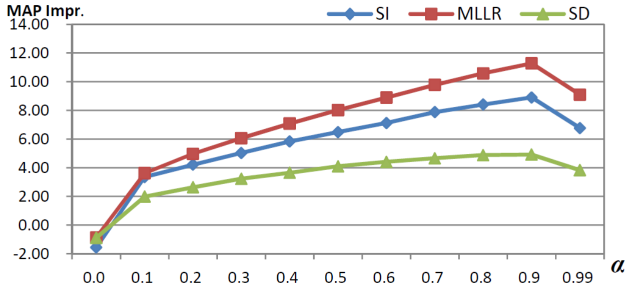

Fig. 4. *Performance improvement for Graph-based Modified Random Walk (G2) compared to first-pass results for values of α in a wide range for the three sets of acoustic models.*

4.2.2. *Graph-Based Re-Ranking*

The results for the two graph-based re-ranking approaches are proposed here, with random walk (G1) as presented in section 3.1 and modified random walk (G2) in section 3.2 are respectively listed in rows (d)(e). The results show that the two graph-based re-ranking methods were significantly better than the first-pass results for all acoustic models, especially when the acoustic models were relatively poorer (SI and MLLR), or the original relevance scores were less precise. They also clearly outperformed the pseudo-relevance feedback approaches (PRF1 and PRF2) in rows (b)(c). This verified the global similarity considered by the graph-based approaches is really useful. Moreover, the modified random walk (G2) is better than random walk (G1). The reason why G2 was better than G1 is probably that for G2 the score contribution from neighboring nodes were based on outgoing edges (A_i) as in (8), exactly matched to the way the edge weights were normalized (also based on A_i) as in (3). However, for G1 (4) was based on B_i; thus slightly mismatched.

4.2.3. *Cascade Approach*

We then cascaded the better approach of pseudo-relevance feedback (PRF2) with the proposed graph-based approaches (G1 and G2). We applied PRF2 first and then on the retrieved set of PRF2 performed the graph-based re-ranking (G1 or G2), and we re-ranked the utterances according to the final scores. The results are listed in rows (f)(g). We see the performance of PRF2 + G2, cascade of the two better approaches, was always better than PRF2 or G2 individually for the relatively poorer acoustic models (SI and MLLR). Thus, the two approaches are clearly additive. The results of PRF2 + G2 were not better for SD model, probably because the local similarity the pseudo-relevance feedback considers was already not far from global optimum, and the additional graph-based re-ranking thus simply perturbed the results.

It is reasonable that the proposed graph-based approach and the previous approach of pseudo-relevance feedback approach are additive, especially for mismatched acoustic models. The former considers the global structure of similarities among all utterances over the graph, while the latter considers primarily the local similarity between the pseudo-relevant utterance set Y_Q and each retrieved utterance.

4.3. Performance Sensibility with α

The results in Table 1 are for $\alpha = 0.9$. It is important to analyze the dependence of the performance on the choice of the value of α. The achievable improvements in MAP compared to row (a) in Table 1 for the better graph-based approach (G2) as in row (e) and its cascade with PRF2 (PRF2 + G2) as in row (g), except with different values of α, are plotted respectively in Fig. 4 and Fig. 5 for the three sets of acoustic models. We can see from the figures that the achievable improvements were relatively stable for a wide range of values of α, and the improvements were maximized when $\alpha = 0.9$ in Fig. 4 and $\alpha = 0.9, 0.8, 0.7$ respectively for SI, MLLR, SD models in Fig. 5. From (4) and (8) such values of α close to 1 indicate that better retrieval relies primarily on global similarity (weighted $0.7 - 0.9$), and the original relevance scores in (1) are really not reliable (weighted $0.1 - 0.3$). This also explains why significant improvements were achieved with the proposed approach.

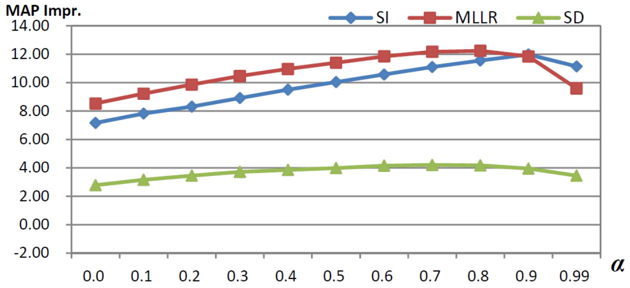

Fig. 5. *Performance improvement for Cascade 2 (PRF2 + G2) compared to first-pass results for values of α in a wide range for the three sets of acoustic models.*

Note that in Fig. 4 the performance was optimized at $\alpha = 0.9$ for all three models. Thus, the graphic structure provided significant information to improve the ranking. However, in Fig. 5, the trends were slightly different. With better acoustic models the pseudo-relevant utterance set Y_Q was easier to select; therefore the original scores were more reliable, and the best value of α was smaller. The results here also show that the proposed approaches are especially useful for mismatched models, which is a highly desired property.

5. CONCLUSIONS

In this paper, we propose graph-based approaches to re-rank the first-pass retrieved utterances to improve the performance of spoken term detection by representing the feature-space similarities as a graph and considering the global structure of these utterances over the graph. Very encouraging results were obtained in the experiments.

6. REFERENCES

[1] R. Rose, A. Norouzian, and A. Reddy, "Subword-based spoken term detection audio course lectures," in *ICASSP*, 2010.

[2] P. Motlicek and F. Valente, "Application of out-of-language detection to spoken term detection," in *ICASSP*, 2010.

[3] D. Wang and S. King, "Stochastic pronunciation modeling and soft match for out-of-vocabulary spoken term detection," in *ICASSP*, 2010.

[4] T. Mertens and D.Schneider, "Merging search spaces for subword spoken term detection," in *InterSpeech*, 2009.

[5] C. Chen and H. Lee, "Improved spoken term detection by feature space pseudo-relevance feedback," in *InterSpeech*, 2010.

[6] C. Parada and A. Sethy, "Query-by example spoken term detection for OOV terms," in *ASRU*, 2009.

[7] L. Page and et al, "The pagerank citation ranking: bringing order to the web," in *Technical Report, Stanford Digital Library Technologies Project*, 1998.

[8] W. Hsu and L. Kennedy, "Video search reranking through random walk over document-level context graph," in *MM*, 2007.

[9] C. Chan and L. Lee, "Unsupervised Spoken-Term Detection with Spoken Queries Using Segment-based Dynamic Time Warping," in *InterSpeech*, 2010.

[10] A. Langville and C. Meyer, "A survey of eigenvector methods for web information retrieval," in *SIAM Review*, 2005.