



Improved Spoken Term Detection by Feature Space Pseudo-Relevance Feedback

Chia-ping Chen, Hung-yi Lee, Ching-feng Yeh, Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University

{coward7652,tlkagkb93901106}@yahoo.com.tw, r98942056@ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

In this paper, we propose an improved approach for spoken term detection using pseudo-relevance feedback. To remedy the problem of unmatched acoustic models with respect to spoken utterances produced under different acoustic conditions, which may give relatively poor recognition output, we integrate the relevance scores derived from the lattices with the DTW distances derived from the feature space of MFCC parameters or phonetic posteriorgrams. These DTW distances are evaluated for a carefully selected set of pseudo-relevant utterances, which obtained from the first-pass returned list given by the search engine. The utterances on the first-pass returned list are then reranked accordingly and finally shown to the user. Very encouraging, performance improvements were obtained in the preliminary experiments, especially when the acoustic models are poorly matched to the spoken utterances.

Index Terms: spoken term detection, pseudo-relevance feedback

1. Introduction

Spoken term detection is to return a list of spoken utterances containing the term requested by the user. In many approaches of spoken term detection, the spoken utterances are first recognized and transformed into transcriptions or lattices by speech recognition technologies, and then the search engine looks through all the transcriptions or lattices very similar to the text-based information retrieval. In this process much of the information in the acoustic signals may be lost in the stage of speech recognition, especially when the acoustic models used are not well matched to the characteristics of the acoustic signals, which naturally results in degraded recognition accuracy and poor detection performance. This is very common in the scenario of spoken term detection, because the huge quantities of spoken utterances available over the Internet are naturally produced by many different people under many different acoustic conditions, it is thus very difficult to train a set of acoustic models well matched to so many different acoustic conditions. As a result, when the relevance scores such as the posterior probabilities of the query term derived from transcriptions or lattices are used to rank the retrieved utterances, it is hard to judge whether a word hypothesis of the query in the transcriptions or lattices is a positive target or a false alarm when the recognition output is unreliable. Although many efficient approaches [1, 2, 3] have been proposed to enhance the detection performance due to the relatively poor recognition output, the compensative information straightly from the feature space is necessary.

In text-based information retrieval, even if the texts to be retrieved include all precise words, it is still difficult to retrieve

all documents relevant to the query term because many of them do not include the very short query term entered by the user. However, because many related terms may co-occur in many related documents, a document containing some words appearing in some documents identified to be relevant by the search engine may have high probability to be relevant, even if it does not include the query term. For example, a document including the words "George Bush", "US", "Middle East" may be relevant to a query term of "White House", even if it does not include the query term of "White House". In other words, it is possible to retrieve the relevant documents without the query term since they are "similar" to some retrieved relevant documents in some way. Pseudo-relevance feedback, also known as blind relevance feedback, is one way to realize the above idea. In this approach, it is assumed that the set of documents appearing on the top of the retrieved document list are relevant (or "pseudo-relevant"), so documents somehow similar to those "pseudo-relevant" documents can be retrieved, for example, by expanding the query with keywords from those "pseudo-relevant" documents [4]. Similar idea of pseudo-relevance feedback has been applied on spoken term detection [5].

In this paper, we try to perform similar pseudo-relevance feedback for spoken term detection as shown in Figure 1. The upper half of Figure 1 is the conventional spoken term detection. MFCC features were obtained from all spoken utterances in the archive, speech recognition produces lattices for the utterances, and the retrieval engine selects the utterances based on the relevance scores evaluated from the lattices with respect to the query Q entered by the user. The approach proposed here in this paper is shown in the lower half of Figure 1. The first-pass returned list is not shown to the user, but instead a "pseudo-relevant utterance set \mathcal{X}_Q " is selected from it, as in the lower half corner of Figure 1. Two approaches for this selection are proposed, direct selection and integrated selection, as in the lower middle block of Figure 1. Similarity between each utterance x on the first-pass returned list and all utterances in the selected set \mathcal{X}_Q is then evaluated using DTW distances derived from either MFCC parameters or phonetic posteriorgram from the recognition output. This similarity is integrated with the original relevance scores obtained from the lattices, and the integrated scores are used to rerank all the utterances in the first-pass returned list. The reranked results are finally shown to the user.

Below, Section 2 presents the details of the proposed approach. Section 3 describes the experiments and discusses the results. The concluding remarks are finally given in Section 4.

2. Proposed Method

The proposed approach is shown in Figure 1. First, the conventional spoken term detection ranks the utterances x based on the

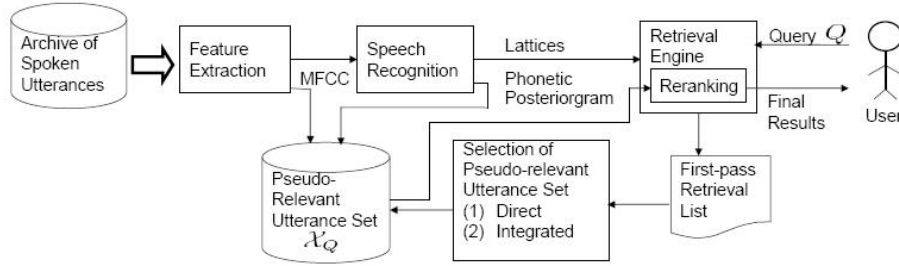


Figure 1: The complete framework of the proposed approach

relevance scores $S_Q(x)$ with respect to the entered query Q derived from the lattices at the speech recognition output as shown in the upper half of Figure 1. This is presented in Section 2.1. Some utterances on the first-pass returned list are then selected to form a pseudo-relevant utterance set as in Section 2.2. The similarity between each utterance x on the first-pass returned list and the pseudo-relevant utterance set \mathcal{X}_Q are then evaluated as presented in Section 2.3. In Section 2.4, all utterances x on the first-pass returned list are reranked based on the integration of the original relevance scores and the similarity with the pseudo-relevant utterances obtained in Section 2.3. Section 2.5 presents two approaches to select the pseudo-relevant utterance set, the direct selection and the integrated selection. Section 2.6 describes how this feedback can be performed iteratively.

2.1. Relevance scores derived from the lattices

The audio signal is divided into utterances, and then each utterance is transcribed into a lattice. The relevance score $S_Q(x)$ of an utterance x with respect to the query term Q is defined as:

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

where a is any arc in the corresponding lattice of x , $word(a)$ is the word hypothesis of a and $P(a|x)$ is the posterior probability. Similar relevance score is widely used in many spoken term detection techniques [6, 7]. After computing $S_Q(x)$ of all utterances x , the retrieval engine selects those utterances x including the word hypothesis Q in their lattices and ranks them by $S_Q(x)$. This gives the first-pass returned list, and the ranking results at this stage are not shown to the user. Here, we assume the query Q is composed of only a single word for simplicity, although the extension to multi-word queries is trivial.

2.2. Construction of pseudo-relevant utterance set selection

Similarity to the conventional pseudo-relevance feedback methods in text retrieval, N utterances on the first-pass returned list, usually those on the top of the list, are taken and assumed to be relevant due to their high relevance scores. These utterances form a pseudo-relevant utterance set $\mathcal{X}_Q = \{x_1, x_2, \dots, x_N\}$ with respect to the query term Q .

2.3. Distance evaluation

Having \mathcal{X}_Q collected, we then compute the similarity between the whole set \mathcal{X}_Q and each utterance x in the first-pass returned list based on their "hit regions". The "hit region" is defined in Figure 2, which is 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

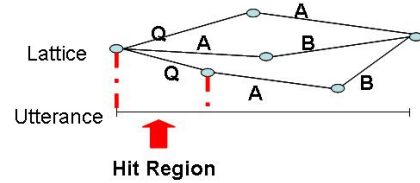


Figure 2: The hit region is defined as the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query term Q with the highest posterior probability in the lattice.

distance between two hit regions, one in the utterance x and the other in an utterance in \mathcal{X}_Q , both represented as sequences of vectors (MFCC vectors or phonetic posteriorgram vector in this research), is then computed with Dynamic Time Warping (DTW) [8]. The distance between each utterance x in the first-pass returned list and the whole pseudo-relevant utterance set \mathcal{X}_Q is then

$$D(\mathcal{X}_Q, x) = \sum_{i=1}^N d(x_i, x)^2, \quad (2)$$

where $d(x, y)$ is the DTW distance between two vector sequences for the two hit regions of utterances x and y , and the summation in equation (2) is over all utterances in the pseudo-relevant utterance set, $\{x_1, x_2, \dots, x_N\}$. So, $D(\mathcal{X}_Q, x)$ is the distance between each utterance x and the whole set \mathcal{X}_Q . In performing DTW, Euclidean distance is used for MFCC vectors, while negative logarithm of inner product is used for phonetic posteriorgram vectors [9].

2.4. Score integration

We transform the distance $D(\mathcal{X}_Q, x)$ obtained above into a similarity measure between \mathcal{X}_Q and x :

$$SIM(\mathcal{X}_Q, x) = 1 - D(\mathcal{X}_Q, x)/M_Q, \quad (3)$$

where M_Q is the largest value of $D(\mathcal{X}_Q, x)$ for all utterances x in the first-pass returned list for the query Q . With this definition, we simply normalize the score into a range between zero and one indicating the similarity. Finally, we integrate the similarity with the original relevance score as below,

$$\hat{S}_Q(x) = S_Q(x)(SIM(\mathcal{X}_Q, x))^\delta, \quad (4)$$

where δ is a weighting parameter. We can then rerank all the utterances in the first-pass returned list using this integrated relevance score $\hat{S}_Q(x)$.

2.5. Selection of the pseudo-relevant utterance set

The N utterances, $\{x_1, x_2, \dots, x_N\}$, in the pseudo-relevant utterance set \mathcal{X}_Q is usually assumed to be the top N utterances on the first-pass returned list with the highest relevance scores $S_Q(x)$ in equation (1). However, with relatively poor acoustic models, it is unavoidable to include some completely irrelevant utterances in this top N utterance set due to recognition errors. To better handle this problem, we set a threshold λ for the relevance score $S_Q(x)$, collect a reference utterance set $\hat{\mathcal{X}}^{Q,R}$ for all utterances with $S_Q(x)$ exceeding λ , and calculate a score for each utterance x in the first-pass returned list with respect to this reference utterance set $\hat{\mathcal{X}}^{Q,R}$.

$$C_Q(x) = S_Q(x) + \gamma \text{SIM}(\hat{\mathcal{X}}^{Q,R}, x), \quad (5)$$

where $S_Q(x)$ is in equation (1), $\text{SIM}(\hat{\mathcal{X}}^{Q,R}, x)$ is evaluated as in equation (2)(3), and γ is a weighting parameter. We can then select N utterances x using this score $C_Q(x)$ in equation (5) from the first-pass returned list to form a better pseudo-relevant utterance set \mathcal{X}_Q . In this way, all utterances in this set \mathcal{X}_Q not only has high relevance score $S_Q(x)$, but is similar enough to all utterances with high $S_Q(x)$ based on DTW distances between hit regions. In this way, the feature-based similarity and model-based relevance scores are both considered in the selection of the pseudo-relevant utterance set to make sure only relevant enough utterances are included. In this paper, the conventional approach of taking N utterances with highest score $S_Q(x)$ to form the pseudo-relevant utterance set is referred to direct selection, while the approach proposed here to select N utterances with highest $C_Q(x)$ in equation (5) is referred to integrated selection.

2.6. Recursive iterations

The above procedure of pseudo-relevance feedback can be performed iteratively. Similar approach has been proposed in text-base information retrieval [10]. Equation (4) at the i^{th} iteration can be written as

$$\hat{S}_Q^i(x) = S_Q^i(x) (\text{SIM}(\mathcal{X}_Q^i, x))^\delta, \quad (6)$$

where $S_Q^i(x)$ at the i^{th} iteration is $\hat{S}_Q^{i-1}(x)$ obtained at the $(i-1)^{\text{th}}$ iteration, and the pseudo-relevant utterance set \mathcal{X}_Q^i at the i^{th} iteration is selected based on the new score $S_Q^i(x)$ obtained at the i^{th} iteration.

3. Experiments

3.1. Experimental setup

In our experiments, the corpus used was the recorded lectures of a course offered in National Taiwan University, including a total of 45 hours of speech produced by a single instructor. The speech was spontaneous and relatively noisy, primarily in the host language of Mandarin Chinese, but with all terminologies produced in the guest language of English embedded in the Mandarin speech. Such code-switching environment is very common for the lectures offered in Taiwan, but makes the recognition task more difficult because the acoustic models, lexicon and language model all need to consider the code-switching environment. 12 hours of the corpus were used in training the acoustic models, while the other 33 hours used as the archive of speech utterances to be retrieved.

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 the 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 has 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. Furthermore, we selected 162 single-word text queries as the testing query set, and the mean average precision (MAP) was used to evaluate the performance of the proposed approach. In our experiment, the baseline is the MAP for retrieved results simply ranked by the relevance scores in equation (1) derived from the lattices.

3.2. Direct selection of \mathcal{X}_Q

In this section, the pseudo-relevant utterance set \mathcal{X}_Q is chosen by direct selection of the top N utterances in the returned list with the highest $S_Q(x)$, and equation (4) is used to rerank the utterances in the first-pass returned list. The number of selected utterances N varied from 1 to 15. Table 1 lists the results when

Table 1: The results for direct selection of \mathcal{X}_Q and integrating the DTW distances from MFCC and phonetic posteriorgram, respectively for three sets of acoustic models, with the number of pseudo-relevant utterances N ranging from 1 to 15. Max RI is the maximal relative improvement achieved for a certain N .

MAP	MFCC			Posteriorgram		
	SI	MLLR	SD	SI	MLLR	SD
baseline	45.47	55.54	73.52	45.47	55.54	73.52
N=1	48.98	56.65	74.01	44.02	53.86	74.04
N=3	48.14	58.75	75.33	44.45	53.95	74.76
N=5	49.32	59.94	75.55	44.71	54.01	75.04
N=7	49.60	59.89	75.12	44.85	54.10	75.13
N=9	49.02	59.65	75.08	44.90	54.13	75.17
N=11	48.91	59.80	74.60	44.95	54.16	75.14
N=13	48.64	59.45	74.30	44.98	54.17	75.09
N=15	48.45	59.03	74.23	44.99	54.16	75.05
Max RI(%)	9.08	7.92	2.76	-	-	2.24

MFCC and phonetic posteriorgram were used in evaluation of DTW distances in equation (2). The maximal relative improvement (Max RI) achieved in the best case for a certain N as compared to the baseline in the first row is listed in the last row of the table. From Table 1, we can see the integration of the DTW distance from MFCC and relevance score from the recognition output outperformed the baseline for all the three sets of acoustic models, while that using DTW distances from phonetic posteriorgram could not do better than the baseline with SI and MLLR acoustic models, although with some improvements with SD models. Because the phonetic posteriorgram was obtained based on the recognition output which included inevitable recognition errors, very similar to the relevance scores obtained from the lattices, it is reasonable that

not too much extra information can be brought here by the integration, especially with relatively poor acoustic models and relatively poor recognition accuracy. Some observations can be made for the results of using DTW distances from MFCC. First, the maximal relative improvement (Max RI) is higher for less matched acoustic models, clearly because the distances and similarity evaluated in the feature space of MFCC is complementary to the relevance scores evaluated from the recognition output which may include many recognition errors. The integration of those two different information sources therefore helped. So the distances and similarity evaluated in the feature space of MFCC can improve the retrieval performance when the recognition accuracy is relatively low, which is a very good desired property. Second, in most cases, MAP first increased to a peak and then decreased as N was raised. This is reasonable because we had more different relevant examples when N increased and the influence of incorrectly selected irrelevant utterances was diluted, but when N was too large, more noisy or possibly irrelevant utterances were selected, which naturally degraded the MAP. This implied the proper selection of N makes sense.

3.3. Integrated selection of \mathcal{X}_Q

Table 2: The results for integrated selection of \mathcal{X}_Q for the three different sets of acoustic models, and different number of pseudo-relevant utterances N . Max RI is the maximal relative improvement.

MAP	SI	MLLR	SD
baseline	45.47	55.54	73.52
N=1	49.54	59.83	74.04
N=5	49.74	60.63	74.96
N=9	49.81	60.75	75.12
N=13	49.75	60.63	75.09
Max RI(%)	9.54	9.38	2.18

The results of using integrated selection of the pseudo-relevant utterance set \mathcal{X}_Q based on the score of equation (5) are listed in Table 2. Only DTW distances from MFCC were used here, since MFCC offered improved performance in Table 1 but phonetic posteriorgram did not. We can see that better performance was achieved by the integrated selection of equation (5) as compared to the direct selection in Table 1 for both the SI and MLLR models with relatively poor recognition accuracy. However, no improvement was obtained with the better matched SD model which gave highly reliable ranking for the first-pass returned list. Also, we find that the MAP results here in Table 2 are much less dependent on the number of pseudo-relevant utterances N chosen as compared to Table 1, and the performance of integrated selection in Table 2 was significantly better than those in Table 1 when the number of N is especially large (e.g. $N = 13$) and small (e.g. $N = 1$), because this integrated selection approach was able to select better pseudo-relevant utterances while reject irrelevant utterances. So the integrated selection approach is more robust to the varying quality of the acoustic models as well as the number of selected pseudo-relevant utterances.

3.4. Recursive iterations

We performed the above pseudo-relevance feedback (with direct selection as in Table 1 and integrated selection as in Table 2) iteratively to see if further improvements were achievable.

Only DTW distances from MFCC were used. N is set to 5 here, and except for the first-iteration, the weighting factor δ in equation (6) was set to 1 to avoid over-fitting for the new scores. The results are listed in Table 3. The middle row of 1 iteration corresponds to those in Table 1 and 2. We observe that further improvements were achieved with more iterations.

Table 3: The results when more iterations were performed for both direct selection and integrated selection of \mathcal{X}_Q . The number of pseudo-relevant utterances N was 5. RI is the relative improvements as compared to the baseline.

MAP	Direct Selection			Integrated Selection		
	SI	MLLR	SD	SI	MLLR	SD
baseline	45.47	55.54	73.52	45.47	55.54	73.52
iter 1	49.32	59.94	75.55	49.74	60.63	74.96
iter 5	49.87	60.26	75.67	50.44	61.43	75.17
RI(%)	9.68	8.50	2.92	10.93	10.60	2.24

4. Concluding Remarks

In this paper, we propose to feedback the feature space information of the pseudo-relevant utterances obtained in the first-pass returned list to improve the performance of spoken term detection. It was formed that the integration of the feature space information derived from the hit regions of the relevant utterances with the relevance scores derived from recognition output can actually improve the detection performance, especially with acoustic models under different matched conditions. We also verify that MFCC parameters offered very useful feature space information here. Both integrated selection of the pseudo-relevant utterances and recursive iterations of feedback are formed to be useful too.

5. References

- [1] R. Wallace, R. Vogt, B. Baker, and S. Sridharan, "Optimising figure of merit for phonetic spoken term detection," in *ICASSP*, 2010.
- [2] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," in *ICASSP*, 2010.
- [3] D. Wang, S. King, J. Frankel, and P. Bell, "Term-dependent confidence for out-of-vocabulary term detection," in *Interspeech*, 2009.
- [4] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving pseudo-relevance feedback in web information retrieval using web page segmentation," in *International World Wide Web Conference*, 2003.
- [5] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for oov terms," in *ASRU*, 2009.
- [6] H.-Y. Lee and L.-S. Lee, "Integrating recognition and retrieval with user feedback: A new framework for spoken term detection," in *ICASSP*, 2010.
- [7] Z.-Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken document retrieval for the internet: lattice indexing for large-scale web-search architectures," in *HLT NAACL*, 2006.
- [8] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition," in *ICSLP*, 2006.
- [9] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009.
- [10] O. Kurland, L. Lee, and C. Domshlak, "Better than the real thing?: iterative pseudo-query processing using cluster-based language models," in *SIGIR*, 2005.