TOWARDS UNSUPERVISED SEMANTIC RETRIEVAL OF SPOKEN CONTENT WITH QUERY EXPANSION BASED ON AUTOMATICALLY DISCOVERED ACOUSTIC PATTERNS

Yun-Chiao Li [1], *Hung-yi Lee* [2], *Cheng-Tao Chung* [3], *Chun-an Chan* [1], *and Lin-shan Lee* [1]

[1]Graduate Institute of Communication Engineering, National Taiwan University
[2]Research Center for Information Technology Innovation, Academia Sinica
[3]Graduate Institute of Electrical Engineering, National Taiwan University

(ychiaoli18, tlkagkb93901106, b97901182, chunanchan)@gmail.com, lslee@gate.sinica.edu.tw

ABSTRACT

This paper presents an initial effort to retrieve semantically related spoken content in a completely unsupervised way. Unsupervised approaches of spoken content retrieval is attractive because the need for annotated data reasonably matched to the spoken content for training acoustic and language models can be bypassed. However, almost all such unsupervised approaches focus on spoken term detection, or returning the spoken segments containing the query, using either template matching techniques such as dynamic time warping (DTW) or model-based approaches. However, users usually prefer to retrieve all objects semantically related to the query, but not necessarily including the query terms.

This paper proposes a different approach. We transcribe the spoken segments in the archive to be retrieved through into sequences of acoustic patterns automatically discovered in an unsupervised method. For an input query in spoken form, the top-N spoken segments from the archive obtained with the first-pass retrieval with DTW are taken as pseudo-relevant. The acoustic patterns frequently occurring in these segments are therefore considered as query-related and used for query expansion. Preliminary experiments performed on Mandarin broadcast news offered very encouraging results.

Index Terms— Query by Example, Query Expansion, Semantic Retrieval

1. INTRODUCTION

Unsupervised spoken content retrieval with spoken query has become popular recently, in which the difficulties of obtaining annotated corpora reasonably matched to the spoken content for training acoustic and language models for speech recognition is bypassed. With the user query entered in spoken form, the similarity between the query and the spoken content can be computed via template matching techniques such as dynamic time warping (DTW) [1, 2]. Therefore, no supervised model training is needed. Since DTW is limited in modeling signal variations, posteriorgrams [1] and acoustic segment models were used to incorporate the signal variation and temporal information [3, 4]. Furthermore, in recent years unsupervised discovery of acoustic patterns has become successful in recent years and such patterns have been shown useful for spoken term detection [5, 6, 7, 8].

Although the above series of approaches have bypassed the difficulties of speech recognition, what they are primarily able to accomplish is detection of the spoken terms, or to return the spoken segments containing the query terms. However, the user prefers to retrieve all objects semantically related to the query regardless of whether the query is included or not. For example, when the query "U.S. president" is entered, the spoken segments including the term "White House" should be returned even if they do not include "U.S. president". However, unsupervised approaches for such semantic retrieval of spoken content without using speech recognition has not been reported yet.

Retrieval of semantically related spoken content [9, 10] has been investigated [11], but all previous works for such purposes utilized speech recognition to transcribe the spoken archive in order to analyze the semantic relationships. Taking the ASR transcriptions as the text, query expansion techniques developed for text information retrieval can be directly applied for such purposes, in which words are semantically correlated to the query can be automatically identified and added to the query [12]. An example approach is to utilize the pseudo relevance feedback (PRF) concept. In this approach the top-N segments in the first-pass retrieval results are assumed to be pseudo-relevant. Those words frequently occurring in these pseudo relevant segments are then used to expand the query, and therefore the spoken segments not containing the query term can be retrieved. All these semantic retrieval approaches for spoken content rely on reasonable quality of the ASR transcriptions of the spoken content, which is practically difficult in many situations.

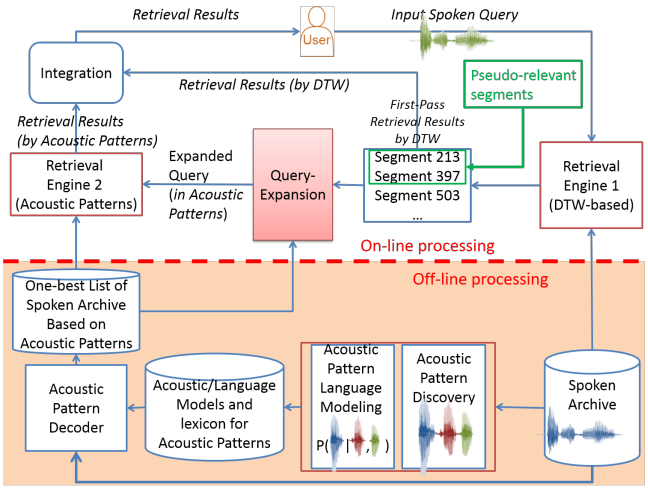


Fig. 1. The framework of the proposed approach.

This paper presents the first known effort of semantic retrieval

of spoken content without using any annotated corpora, but based on query expansion with automatically discovered acoustic patterns obtained in a completely unsupervised way. The framework of the proposed approach is shown in Fig. 1. At the lower half of the figure for off-line processing, acoustic patterns including acoustic/language models and lexicon for them are automatically discovered from the spoken archive in an unsupervised manner. With these acoustic/language models and lexicon for the acoustic patterns, an Acoustic Pattern Decoder at the lower left corner at Fig. 1 (very similar to an ASR decoder but it is not ASR since it is for acoustic patterns, not for phonemes or words) is constructed and performed to transcribe each spoken segment in the archive into a one-best list. On the upper half of Fig. 1 for on-line processing, when a spoken query is entered by the user, the conventional DTW in Retrieval Engine 1 on the right is applied to compute the similarity between the spoken segments and the query to generate the first-pass results on the middle right, in which the top-ranked segments are taken as pseudo-relevant. Those acoustic patterns frequently occurring in these pseudo-relevant segments are thus assumed to correspond to the terms semantically related to the query and used in the Query Expansion in the middle. Therefore the Retrieval Engine 2 based on acoustic patterns on the left can search for the spoken segments containing those query-related patterns. The results of Retrieval Engine 2 are finally integrated with the DTW results from Retrieval Engine 1 and then presented to the user.

2. PROPOSED APPROACH

2.1. Preprocessing - Acoustic Pattern Discovery

Unsupervised signal pattern discovery techniques [13, 14, 15, 16, 17] have been widely developed to identify repeated acoustic patterns. Such techniques have been utilized for enhancing spoken document classification [18, 19], spoken term detection [5, 6, 7, 8], music retrieval [20], video retrieval [21] and spoken document retrieval [22]; but not yet properly leveraged for semantic retrieval of spoken content. Here in this study the recently proposed approach [5, 22] for discovering the two-level acoustic patterns is employed, which includes subword-like and word-like patterns (a word-like pattern is a sequence of one to several subword-like patterns), the lexicon of word-like patterns in terms of subword-like patterns, and the n-gram language model for word-like patterns. Each subword-like acoustic patterns is modeled as an HMM. All parameters including HMM parameters for the subword-like acoustic patterns, the alphabet size of subword-like patterns, the lexicon size of word-like patterns, and the n-gram language model parameters for the word-like patterns are all automatically learned in an unsupervised manner [5] from the spoken archive to be retrieved. This is achieved by integrating a dynamic lexicon into the process of the conventional HMM-training, and performing three stages of iterative optimization between the assumed labels and the trained models, such that the models, parameters, and the two-level linguistic structure can then collect knowledge from the corpora layer after layer iteratively and adjust themselves accordingly [5]. These acoustic/language models and lexicon for two-level acoustic patterns are used to construct the Acoustic Pattern Decoder completely based on these acoustic patterns (very similar to but not an ASR decoder, with word-like patterns considered like words and subword-like patterns like phonemes), which generates a one-best list in terms of word-like acoustic patterns for each spoken segment. This is shown on the left of the lower part of Fig. 1. These one-best lists in terms of word-like acoustic patterns will be utilized for query expansion in Sections 2.3- 2.5.

2.2. Retrieval Engine 1 - Frame-based DTW

In the Retrieval Engine 1 on the right of the upper part of Fig. 1, frame-based DTW is performed for the input spoken query against all segments in the spoken archive, in order to retrieve the spoken segments containing the terms in the spoken queries. In the DTW process the two feature vector sequences of different lengths, one for the spoken segment and the other for the query, are matched for distance evaluation based on an optimal warping path [2]. The segments in the spoken archive with minimum distances to the query are obtained as the first-pass retrieved results ranked by the distances. Those spoken segments ranked top N on the list are taken as pseudo relevant.

2.3. Language Model Retrieval Approach and Query Expansion

The purpose of Retrieval Engine 2 on the left of upper part of Fig. 1 is to retrieve spoken segments semantically related to the query but not necessarily including the query terms. This can be achieved by the query expansion approach based on the language modeling retrieval approach [23], but here this approach has to be performed with acoustic patterns as described below. The basic idea is that we respectively represent each spoken segment x in the archive and the query Q as language models, θ_x and θ_Q, but these language models are expressed in terms of acoustic patterns obtained off-line in Section 2.1 at the lower part of Fig. 1, rather than in terms of words as in conventional language models. In other words, all language models below refer to probabilities of acoustic patterns instead of words. The KL divergence $KL(\theta_Q|\theta_x)$ between the language models θ_Q and θ_x is used to evaluate the relevance score $S(Q,x)$ between a segment x and the query Q. All language models below are unigram plus bigram plus trigram models in the experiments reported below, although the proposed approach is not limited to this case. Additionally, all language models below refer to probabilities of acoustic patterns instead of words. With pseudo-relevant spoken segments identified by DTW as in Section. 2.2, the acoustic patterns frequently appearing in these pseudo-relevant segments may represent some terms semantically related to the query. Therefore, the counts of such frequently appearing acoustic patterns can be employed to expand the language model for the query. For example, for the query "U.S. president", the acoustic patterns for the term "White House" may occur frequently in the pseudo-relevant segments, thus these patterns can be used to expand the query language model θ_Q. As a result, the expanded query language model θ_Q includes the acoustic patterns not only for terms in the query, but also for those semantically related to the query.

2.4. Acoustic Pattern Language Model for Spoken Segments

Here, we describe how to obtain the segment model θ_x for a spoken segment x. With each spoken segment x in the archive transcribed into a one-best list of the word-like acoustic patterns and then further expressed as a sequence of subword-like acoustic patterns as described in Section 2.1, a language model θ'_x based on the one-best result can be obtained:

$$P(t|\theta'_x) = \frac{C(t,x)}{\sum_t C(t,x)} \quad (1)$$

where t is the label of a unigram, bigram or trigram of the subword-like patterns, and $C(t,x)$ is the count of t in the sequence of subword-like patterns in the one-best decoded result of the segment x. Similarly we can estimate an acoustic pattern background model

θ_b trained from the whole spoken archive:

$$P(t|\theta_b) = \frac{\sum_{x \in C} C(t,x)}{\sum_t \sum_{x \in C} C(t,x)} \quad (2)$$

which is the probability of observing the n-gram with label t for the subword-like acoustic patterns in the whole spoken archive, and C is the spoken archive considered. The segment model θ_x to be used below is then the interpolation of θ'_x in (1) and the background model θ_b in (2):

$$P(t|\theta_x) = \alpha P(t|\theta'_x) + (1-\alpha) P(t|\theta_b) \quad (3)$$

where α is the weight for interpolating $P(t|\theta'_x)$ and $P(t|\theta_b)$.

2.5. Acoustic Pattern Language Model for Expanded Query

Here we adopt the query-regularized mixture model widely used for text information retrieval, but use it with the acoustic patterns. This model assumes that the words in pseudo-relevant documents are either query-related words or general words, with a document-dependent ratio between the two. For example, for those irrelevant documents taken as pseudo-relevant, this ratio for the query-related words to the general ones should be very low. These document-dependent ratios and which words are query-related are actually unknown, but can be estimated from the pseudo-relevant documents. Therefore, query-related word distributions can be estimated with this model from the pseudo-relevant documents, based on which the query language model can be expanded.

The above model is adopted here directly, except the words in the documents are replaced by acoustic patterns in the spoken segments. Therefore, for the work here, distributions for query-related acoustic patterns are estimated from the pseudo-relevant spoken segments obtained from DTW , which are used to construct the expanded query language model θ_Q in terms of acoustic patterns. Suppose the N pseudo-relevant spoken segments (that is, the top N segments in the first-pass retrieval results ranked by DTW) are $x_1, x_2, ..., x_n, ..., x_N$. With the assumption that the acoustic patterns in each pseudo-relevant spoken segment are either query-related or general, the segment language model θ'_{x_n} of (1) for each pseudo-relevant segment x_n should be close to an estimated model θ''_{x_n} which is the interpolation of the expanded query model θ_Q to be estimated (for query-related acoustic patterns) and the background language model θ_b in (2) (for general acoustic patterns) with a segment dependent weight α_n.

$$P(t|\theta''_{x_n}) = \alpha_n P(t|\theta_Q) + (1-\alpha_n) P(t|\theta_b), \quad (4)$$

where α_n is the segment-dependent interpolation weight for segment x_n, which is to be estimated as well. Therefore, we utilized this segment language model in (4) to minimize (5) in order to estimate the expanded query model θ_Q.

$$F_1(\theta_Q, \alpha_1, ..., \alpha_N) = \sum_{n=1}^{N} KL(\theta'_{x_n} | \theta''_{x_n}) \quad (5)$$

which means the query model θ_Q should be expanded in such a way that the sum of the KL divergence between each segment model θ'_{x_n} in (1) and the corresponding interpolated language model θ''_{x_n} in (4) for all the N pseudo-relevant segments is minimized.

However, the expanded query model θ_Q minimizing (5) can be just for the common acoustic pattern distributions in the pseudo-relevant segments, not necessarily query-related. To handle this problem, θ_Q should be regularized by an original query model θ'_Q. However, a mismatched situation exists here. The query entered by the user is in spoken form, only the frame-based DTW template matching was performed between the query and the spoken segments. If the query is simply decoded by the Acoustic Pattern Decoder trained from the spoken archive as described in Section 2.1, then a serious mismatch in signals between the spoken query and the spoken segments in the archive (such as those due to very different speakers, speaking rates and acoustic conditions) may produce very seriously corrupted acoustic patterns significantly different from those corresponding to the terms in the query. This problem can be properly solved by the frame-based DTW performed by Retrieval Engine 1 as described in Section 2.2. DTW provides the hypothesized regions within those top N spoken segments in the first-pass retrieved results which may possibly contain the query. On the other hand, all the spoken segments in the archive have already been decoded into one-best sequences of acoustic patterns using the Acoustic Pattern Decoder completely based on acoustic patterns as shown in the lower left corner of the lower part of Fig. 1 and described in Section 2.1. Therefore, we can align the hypothesized regions for the top N pseudo-relevant segments obtained from DTW to the one-best sequences of acoustic patterns for those segments. In this way a query model θ'_Q representing the original query can be estimated as below:

$$P(t|\theta'_Q) = \frac{\sum_{n=1}^{N} C'(t, x_n)}{\sum_t \sum_{n=1}^{N} C'(t, x_n)} \quad (6)$$

where $C'(t, x_n)$ is the count of acoustic pattern n-gram t appearing in the hypothesized region of a pseudo-relevant spoken segment x_n. Here the acoustic patterns totally within the hypothesized region is counted as one, while those acoustic patterns with only a part of the signal within the hypothesized regions are counted by the percentage of the duration of the acoustic pattern within the hypothesized region. The numerator of (6) is the sum over all N pseudo-relevant segments, and further summed over all acoustic pattern n-grams t in the denominator of (6) for normalization. The goal here is then to have the expanded query model θ_Q not too different from the model θ'_Q in (6) representing the original query, or minimizing (7) below,

$$F_2(\theta_Q, \theta'_Q) = KL(\theta'_Q | \theta_Q) \quad (7)$$

which is the KL divergence between θ'_Q and θ_Q.

The final expanded query model θ_Q is then obtained by minimizing both the KL divergence with all pseudo-relevant segments as in (5) and the KL divergence with the original query as in (7) at the same time. Therefore the expanded query model θ_Q and the corresponding weights α_n for all pseudo-relevant segments in (4) are actually estimated by minimizing the following objective function:

$$F(\theta_Q, \alpha_1, ..., \alpha_N) = F_1(\theta_Q, \alpha_1, ..., \alpha_N) + \mu F_2(\theta_Q, \theta'_Q) \quad (8)$$

where the first term $F_1(\theta_Q, \alpha_1, ..., \alpha_N)$ in (5) is to learn the common pattern distribution from all the pseudo-relevant segments, and the second term $\mu F_2(\theta_Q, \theta'_Q)$ is to make sure θ_Q is close enough to θ'_Q in (6). μ is a parameter controlling the influence of the second term.

2.6. Retrieval Engine 2 and Integration

The Retrieval Engine 2 on the left of the upper part of Fig.1 then simply performs the language modeling retrieval approach by evaluating the KL divergence $KL(\theta_Q|\theta_x)$ between the expanded query model θ_Q obtained by minimizing (8) in Section 2.5 and the segment

model θ_x in (3) of Section 2.4.

The final results shown to the user is then ranked according to the weighted sum $S(Q,x)$ of the two relevance scores, the first $R_{QE}(Q,x)$ evaluated with the language model retrieval approach with query expansion, and the second $R_{DTW}(Q,x)$ with DTW, both normalized between 0 and 1.

$$S(Q,x) = -[w_1 R_{DTW}(Q,x) + (1-w_1) R_{QE}(Q,x)] \quad (9)$$

3. EXPERIMENTS

3.1. Experimental Setup

The spoken archive to be retrieved through in the experiments included 4 hours of Mandarin broadcast news stories collected daily from local radio stations in Taiwan in 2001, and further manually segmented into 5034 spoken segments, each with one to three utterances. We selected 30 spoken terms as the test queries, and all collected from the speakers in the same corpus. All utterances containing the spoken queries were excluded from the 5034 utterances. Two answer sets were used for evaluation. The first answer set was for semantic retrieval, which contained not only those spoken segments including the query, but also those semantically related to the query but not including it in the utterances. The second answer set was for the conventional spoken term detection, which simply contained the spoken segments including the query. The first set contained in average 171.9 relevant segments for each query, and the second in average 27.6 relevant segments for each query. Mean average precision (MAP) was used as the performance measure.

We also performed some supervised approaches utilizing ASR with models trained with annotated data for comparison. For these experiments, a trigram language model trained from a 40M news corpus collected in 1999 and a lexicon of 62K words was used for recognition. The acoustic models included a total of 151 right-context-dependent intra-syllable Initial-Final (I-F) models and it was trained by 8 hrs of broadcast news stories collected in 2000. The recognition character accuracy obtained for the 5034 segments was 75.27%.

3.2. Acoustic Pattern N-grams

With the two-level acoustic patterns and the Acoustic Pattern Decoder as shown in the lower part of Fig. 1, and discussed in Section. 2.1, the acoustic models for the subword-like acoustic patterns, and the language model for the word-like acoustic patterns were trained using the whole corpus of the total 5034 spoken segments. A total of 208 different subword-like acoustic patterns were obtained, and we use the unigram, bigram, and trigram of these subword-like acoustic patterns (a total of 85534) to construct the acoustic pattern language models (θ_x, θ_Q and so on) used in retrieval. Most of the subword-like patterns are very close to syllables in Mandarin, so the bigram and trigram are similar to bi-syllable or tri-syllable words in Mandarin. Some example n-grams of acoustic patterns with the corresponding sample realizations are listed in Table. 1. Row(1) is an example acoustic pattern unigram which sounds similar to the Mandarin syllable /dian/, with sample realizations corresponding to different Chinese mono-syllable characters 店(shop), 點(point), 電(electricity). Row(2) is an example acoustic pattern bigram including the unigram in row(1), with sample realizations corresponding to different Chinese bi-syllable words. For row (3) and (4), it is similar to row (1) and (2). Note that in each case characters or words sounding similarly are clustered together in some acoustic pattern n-grams.

	t (n-grams): (IDs)	Sample Realizations
(1)	unigram: (106)	店(/dian/, shop), 點(/dian/, point), 電(/dian/, electricity)
(2)	bigram: (106)-(27)	電腦(/dian-nau/, computer), 電能(/dian-neng/, electricity)
(3)	unigram: (93)	手(/shou/, hand), 收(/shou/, receive), 熟(/shou/, mature)
(4)	bigram: (93)-(145)	受傷(/shou-shang/, injured), 首相(/shou-shiang/, prime minister)

Table 1. Some examples of unigram and bigram t of the subword-like acoustic patterns and their sample realizations as Chinese mono-syllabic characters and bi-syllabic words

3.3. Experimental Results

The results of the proposed unsupervised approach were listed in the upper part of Table. 2. We list the results of Retrieval Engine 1 or DTW alone ($w_1 = 1.0$ in (9)) in row (1), Retrieval Engine 2 or language modeling retrieval with query expansion alone ($w_1 = 0.0$ in (9)) in row (2), and the integration of them (row(3)) with different weights w_1. Note the Retrieval Engine 2 cannot operate without retrieval engine 1, but in row (2) we simply set $w_1 = 0.0$. Column (A) is for the semantic retrieval discussed here evaluated with the first answer set, while column (B) is for conventional spoken term detection evaluated with the second answer set. The MAP values in column (A) are much lower than those in column (B), obviously because for semantic retrieval the answer set used for column (A) contains much more segments which were semantically relevant but did not include the query, and therefore difficult to identify. Although all values in column (A) are below 10%, these are reasonable for such a very difficult task, just like some other difficult tasks such as video retrieval. It is clear from column (A) that query expansion with automatically discovered acoustic patterns was able to improve the performance of semantic retrieval (row (3) vs. (1) in column (A)), obviously because some semantically relevant segments not containing the query terms may include some acoustic patterns co-occurring with the query terms in some segments detected by DTW. Of course the improvement achieved was not large (e.g. 0.94% in row(4), 9.70% for $w_1 = 0.7$ vs. 8.76% for $w_1 = 1.0$). This implies the initial approaches proposed here may be relatively weak for such a very difficult task and better approaches are definitely needed. It is interesting to note that query expansion with acoustic patterns also improved the performance of spoken term detection (row (3) vs. (1) in column (B)). For spoken term detection, although all relevant segments contain the query term, some of them may not be detected in DTW due to signal mismatch, e.g. speaking rate variation, speaker variation, pronunciation variation or acoustic variation. These query terms may co-occur with some other semantically related terms, so the appearance of some acoustic patterns corresponding to such query-related terms may lead to the detection of these query terms in the spoken segments. The MAP value was not very sensitive to w_1, while $w_1 = 0.7$ seemed to be good, so we set $w_1 = 0.7$ in the following experiments.

Alternatively, we performed supervised semantic retrieval with query expansion using ASR with models trained with annotated data for comparison, and the results are listed in the lower part (row(5)(6)(7)) of Table. 2. First, we use ASR with models trained with annotated data to transcribe the spoken segments into word-based lattices and the queries into one-best word sequences, and then create the word-based unigram language models θ_x and θ_Q

from the lattices using similar approaches as in [22]. Note that the bigram/trigram of subword-like patterns for unsupervised approaches proposed above carry roughly the word-level information, which is in principle in parallel with the word-based unigram model used here. Then the KL divergence between the language models of each spoken segment and each query is computed and used to rank the spoken segments in the first-pass retrieval with results listed in row(5). Then query expansion was performed using the approach similar to that in Section. 2.3 to 2.5, with results listed in row (6). We then integrated the first-pass results with query expansion with $w_1 = 0.7$ exactly as in row(3), with results listed in row (7) and the improvement in row (8). We can see that the supervised approach provided much better performance, but the improvement obtained with query expansion on ASR transcribed lattices and one-best word sequences was not much either, 1.28% (row (8)). This seemed to imply the task considered here was really difficult. While for the unsupervised method, we achieved an improvement of semantic retrieval about 0.94% (row(4)), which is actually comparable to the supervised approach.

	MAP		(A) semantic retrieval	(B) spoken term detection
unsupervised	(1) DTW alone ($w_1 = 1.0$)		8.76%	28.30%
	(2) Query Expansion ($w_1 = 0.0$)		6.03%	7.82%
	(3) Integration	$w_1 = 0.9$	9.28%	29.94%
		$w_1 = 0.7$	9.70%	30.31%
	(4) improvement ($w_1 = 0.7$)		0.94%	2.01%
supervised	(5) first-pass ($w_1 = 1.0$)		29.49%	70.07%
	(6) Query Expansion ($w_1 = 0.0$)		30.30%	68.86%
	(7) Integration ($w_1 = 0.7$)		30.77%	76.80%
	(8) improvement		1.28%	6.73%

Table 2. MAP for semantic retrieval (column(A)) and spoken term detection (column(B)) for DTW alone (row(1)), query expansion alone (row(2)), the integration (row(3)) and the improvement by integration (row(4)), all with the proposed unsupervised approach. Row(5)(6)(7)(8) are for supervised approaches using ASR with models trained with annotated data, including first-pass(row(5)), with query expansion(row(6)), integration(row(7)), and improvements(row(8)). The number of pseudo-relevant segments is $N = 12$, and μ in equation (8) is 300.

In order to analyze how the proposed approach really helped in the goal of semantic retrieval, we collected the top 200 segments ranked by DTW alone and by the proposed approach ($w_1 = 0.7$) given each of the 30 queries. Therefore, total 6000 segments were collected for each approach. In Table. 3, the total numbers of truly semantically relevant segments in the first answer set, with or without query out of the 6000 are listed in column (A). Those segments in column (A) are further divided into two parts, those including the query in column (B) and those not including the query in column (C). Note that in row (1) of Table. 3 for DTW alone, still quite good number (264) of segments not including the query can be detected, probably because most of which included parts of the phoneme sequences of the queries. With the proposed approach ($w_1 = 0.7$) in row (2), we can see that the relevant segments in both columns (B) and (C) are increased, regardless of whether the query is included or not, by roughly $11\% - 15\%$ as shown in row (3).

We further analyze how the results depended on the parameter μ in (8) of Section 2.5, the number N of pseudo-relevant segments

MAP	(A) All semantically relevant segments	(B) Those in (A) including the query	(C) Those in (A) not including the query
(1)DTW alone ($w_1 = 1.0$)	589	325	264
(2) Proposed ($w_1 = 0.7$)	668	374	294
(3) Improved	13.41%	15.07%	11.36%

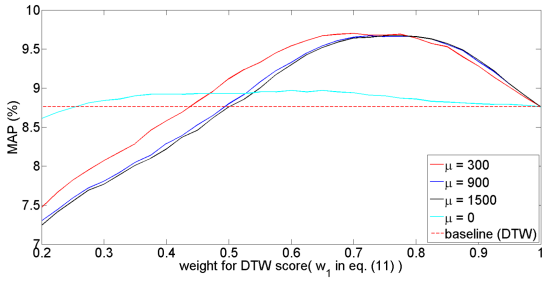
Table 3. Number of all semantically related segments (column(A)), those including (column(B)) and not including (column(C)) the query out of the total of 6000 segments collected from the top 200 segments ranked for the 30 queries, with $N = 12$,$\mu = 300$.

in (5) of Section 2.5, and the weight w_1 in (9) of Section 2.6. This is shown in Fig. 2 (a) and (b). The MAP values with respect to the first answer set for semantic retrieval was plotted as functions of the weight w_1 in the proposed approach integrating DTW and query expansion. In Fig. 2 the baseline is DTW results. In Fig. 2 (a), N is fixed to 12, and results for $\mu = 0, 300, 900, 1500$ are shown. We note the performance was close to the baseline with $\mu = 0$, or the expanded queries were drifted away by the pseudo-relevant segments when $\mu = 0$. The best performance was achieved at $\mu = 300$. We also see improved performance for w_1 ranging from roughly 0.45 to 0.95, or the result was not very sensitive to w_1, although $w_1 = 0.7$ as in Table 2 was a good choice.

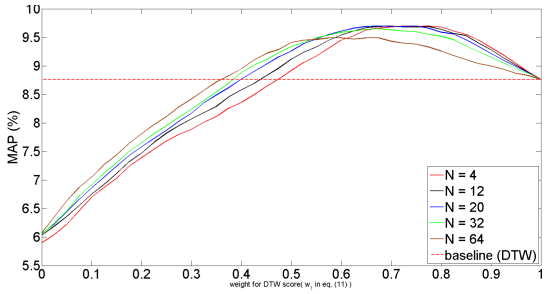
In Fig. 2 (b), μ is fixed to 300, and the results with $N = 4, 12, 20, 32, 64$ are plotted. Similarly we see improved performance for w_1 ranging from 0.45 to 0.95 regardless of the value of N. We note that as N was increased, the MAP first increased a little and then started to decrease when N was larger than 32. This is reasonable, because larger N means that more top-N segments were considered pseudo-relevant, so there were more training data for query expansion. However, if the N was too large, more irrelevant segments were considered relevant, which resulted in performance degradation.

3.4. Semantic Pattern Analysis

Table. 4 lists an example showing that the proposed method indeed expanded the query to include some semantically related acoustic patterns. This table includes the top 5 n-grams with the largest probabilities $P(t|\theta_Q)$ after query expansion for a query "學校(/xue-xiao/, school)". The regions hypothesized to include the query in the top N segments offered by DTW, aligned with the one-best sequences provided by the Acoustic Pattern Decoder, all included the two unigrams listed in rows(1) and (2). Clearly the unigrams in rows(1)(2) were the dominating acoustic patterns for the query. Those n-grams in rows(3)(4)(5) were then added by query expansion. Row(3) is a bigram combined by the original two unigram query in row(1)(2). Row(4) is a unigram pronounced very similar to row(2) but has a different acoustic pattern representation. Row(5) is obviously an added semantically related bigram since it sounds like "學生(/xue-shang/, students)", highly related to the query "學校(/xue-xiao/, school)" but with a very different pronunciation. This verified the query expansion actually worked.



(a) $N = 12$



(b) $\mu = 300$

Fig. 2. MAP yielded by integrating query expansion and DTW. DTW is taken as the baseline. (a) The total number of pseudo-relevant segments $N = 12$, for different values of the weight μ in (8). (b) $\mu = 300$ for different values of N.

	t(n-grams): (IDs)	$P(t\|\theta_Q)$	Sample Realizations
(1)	unigram: (87)	0.4280	校(/xiao/, school)
(2)	unigram: (56)	0.3880	學(/xue/, learning)
(3)	bigram: (56)-(87)	0.0040	學校(/xue-xiao/, school)
(4)	unigram: (129)	0.0030	學(/xue/, learning)
(5)	bigram: (129)-(23)	0.0016	學生(/xue-sheng/, students)

Table 4. Top 5 n-grams in the expanded query θ_Q for the query "學校(/xue-xiao/, school)"

4. CONCLUSIONS

This work presents an initial effort to perform unsupervised semantic retrieval of spoken content using query expansion. Query expansion was originally developed for text retrieval, but here we try to extend it to spoken content with automatically discovered acoustic patterns. We perform query expansion over automatically discovered acoustic patterns using the top N results of DTW. The preliminary experimental results indicate that this approach improves not only the desired semantic retrieval, but the spoken term detection as well, since the co-occurring acoustic patterns also help in identifying the existence of the queries.

5. REFERENCES

[1] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009.

[2] C. a. Chan and L. s. Lee, "Unsupervised spoken-term detection with spoken queries using segment-based dynamic time warping," in *Interspeech*, 2010.

[3] Y. Tsao, H. Sun, H. Li, and C.-H. Lee, "An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition," in *ICASSP*, 2010.

[4] F. K. Soong C.-H. Lee and B.-H. Juang, "A segment model based approach to speech recognition," in *ICASSP*, 1988.

[5] Cheng-Tao Chung, Chan-An Chan, and Lin-Shan Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *ICASSP*, 2013.

[6] Chia-Ying Lee and James Glass, "A nonparametric bayesian approach to acoustic model discovery," in *ACL*, 2012.

[7] Marijn Huijbregts, Mitchell McLaren, and David van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *ICASSP*, 2011.

[8] Haipeng Wang, Cheung-Chi Leung, Tan Lee, Bin Ma, and Haizhou Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *ICASSP*, 2012.

[9] Ciprian Chelba, Timothy J. Hazen, and Murat Saralar, "Retrieval and browsing of spoken content," in *IEEE Signal Processing Magazine*, 2008.

[10] Hung-Yi Lee, Tsung-Hsien Wen, and Lin-Shan Lee, "Improved semantic retrieval of spoken content by language models enhanced with acoustic similarity graph," in *SLT*, 2012.

[11] Tomoyosi Akiba and Koichiro Honda, "Effects of query expansion for spoken document passage retrieval," in *Interspeech*, 2011.

[12] Tao Tao and ChengXiang Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *SIGIR*, 2006.

[13] A.S. Park and J.R. Glass, "Unsupervised pattern discovery in speech," in *Audio, Speech, and Language Processing, IEEE Transactions*, 2008, vol. 16.

[14] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorization," in *Signal Processing Letters, IEEE*, 2008, vol. 15.

[15] Lei Wang, Eng Siong Chng, and Haizhou Li, "An iterative approach to model merging for speech pattern discovery," in *APSIPA*, 2011.

[16] Niklas Vanhainen and Giampiero Salvi, "Word discovery with beta process factor analysis," in *Interspeech*, 2012.

[17] D. F. Harwath, T. J. Hazen, and J. R. Glass, "Zero resource spoken audio corpus analysis," in *ICASSP*, 2013.

[18] Man-Hung Siu, Herbert Gish, Arthur Chan, and William Belfield, "Improved topic classification and keyword discovery using an hmm-based speech recognizer trained without supervision," in *Interspeech*, 2010.

[19] Sourish Chaudhuri, Mark Harvilla, and Bhiksha Raj, "Unsupervised learning of acoustic unit descriptors for audio content representation and classification," in *Interspeech*, 2011.

[20] Matthew Riley, Eric Heinen, and Joydeep Ghosh, "A text retrieval approach to content-based audio retrieval," .

[21] Yang Liu, Wan-Lei Zhao, Chong-Wah Ngo, Chang-Sheng Xu, , and Han-Qing Lu, "Coherent bag of audio words model for efficient large-scale video copy detection," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010.

[22] Hung yi Lee, Yun-Chiao Li, Cheng-Tao Chung, , and Lin-Shan Lee, "Enhancing query expansion for semantic retrieval of spoken content with automatically discovered acoustic patterns," in *ICASSP*, 2013.

[23] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng, "Statistical lattice-based spoken document retrieval," in *ACM Trans. Inf. Syst.*, 2010, vol. 28.