# ENHANCING QUERY EXPANSION FOR SEMANTIC RETRIEVAL OF SPOKEN CONTENT WITH AUTOMATICALLY DISCOVERED ACOUSTIC PATTERNS

*Hung-yi Lee* [#1], *Yun-Chiao Li* [*2], *Cheng-Tao Chung* [†3] *and Lin-shan Lee* [*†4]

Research Center for Information Technology Innovation, Academia Sinica [#]
Graduate Institute of Communication Engineering, National Taiwan University [*]
Graduate Institute of Electrical Engineering, National Taiwan University [†]
tlkagkb93901106@gmail.com[1], ychiaoli18@gmail.com[2], b97901182@gmail.com[3], lslee@gate.sinica.edu.tw[4]

## ABSTRACT

Query expansion techniques were originally developed for text information retrieval in order to retrieve the documents not containing the query terms but semantically related to the query. This is achieved by assuming the terms frequently occurring in the top-ranked documents in the first-pass retrieval results to be query-related and using them to expand the query to do the second-pass retrieval. However, when this approach was used for spoken content retrieval, the inevitable recognition errors and the OOV problems in ASR make it difficult for many query-related terms to be included in the expanded query, and much of the information carried by the speech signal is lost during recognition and not recoverable. In this paper, we propose to use a second ASR engine based on acoustic patterns automatically discovered from the spoken archive used for retrieval. These acoustic patterns are discovered directly based on the signal characteristics, and therefore can compensate for the information lost during recognition to a good extent. When a *text* query is entered, the system generates the first-pass retrieval results based on the transcriptions of the spoken segments obtained via the conventional ASR. The acoustic patterns frequently occurring in the spoken segments ranked on top of the first-pass results are considered as query-related, and the spoken segments containing these query-related acoustic patterns are retrieved. In this way, even though some query-related terms are OOV or incorrectly recognized, the segments including these terms can still be retrieved by acoustic patterns corresponding to these terms. Preliminary experiments performed on Mandarin broadcast news offered very encouraging results.

***Index Terms***— Query Expansion, Acoustic Pattern Discovery

## 1. INTRODUCTION

Spoken content retrieval will be very important to retrieve and browse multimedia content over the Internet. Substantial effort has been made in spoken content retrieval in recent years, and many successful techniques have been developed [1, 2]. Most works in spoken content retrieval nowadays focused on spoken term detection (STD) [3], for which the goal is simply returning spoken segments including the query terms. This is insufficient because users naturally prefer to be offered all objects they are looking for, regardless of whether the query terms are included or not. This leads to substantial recent work on semantic retrieval of spoken content [4, 5, 6, 7, 8, 9, 10, 11, 12]. This paper thus focuses on retrieving semantically related spoken segments using queries in text form.

The core problem of retrieving spoken segments semantically related to the query is that many of such spoken segments may not necessarily contain the query term. A popular approach to this problem in text information retrieval is query expansion, which automatically adds semantically related terms to the query [13, 14, 15]. The expanded query thus can retrieve objects not containing the original query terms but semantically related to the query. Query expansion techniques are very often realized with the concept of pseudo-relevance feedback (PRF). In these methods, the first-pass retrieval results are first generated, in which a small number of top-ranked objects is assumed relevant (or pseudo-relevant). Since many pseudo-relevant objects may include some specific terms semantically related to the query, the original query can be expanded with these terms to retrieve more relevant objects. Taking ASR transcriptions as the text, such query expansion techniques for text information retrieval can be directly applied for spoken content retrieval [4, 5, 7, 9].

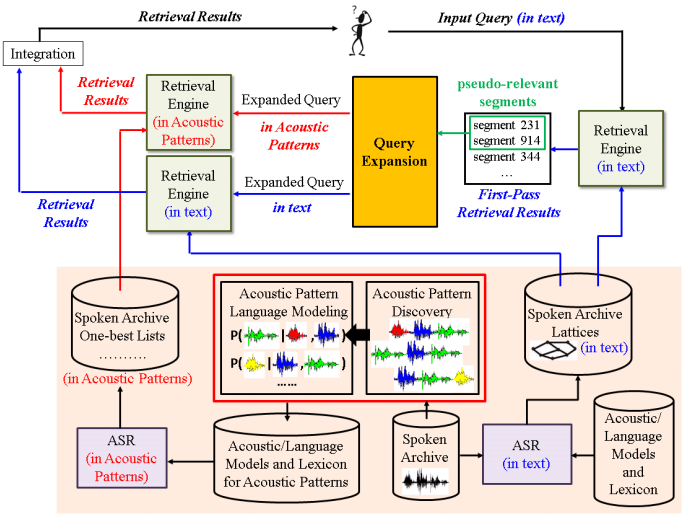


**Fig. 1**. *The framework of the proposed approach.*

However, even if the pseudo-relevant spoken segments actually contain some terms suitable for query expansion, these terms may be OOV or incorrectly recognized, never included in the transcriptions, and therefore cannot help in query expansion. Such OOV problem and recognition errors in ASR lead to the inevitable degradation in the effectiveness of such query expansion techniques when applied on spoken content. In fact, when transcribing speech signals into text, much information is lost and not recoverable, and the OOV problem and recognition errors are just prominent examples. Substantial efforts have been made to try to utilize information carried in the speech signals in spoken content retrieval with a hope to com-

pensate for the loss during recognition [16, 17, 18, 19, 20]. This is also the direction of this paper.

In this paper, we propose to enhance the query expansion for semantic retrieval of spoken content by utilizing a set of acoustic patterns obtained via directly considering speech signal characteristics. The framework of the proposed approach is shown in Fig. 1. At the bottom of Fig. 1, each spoken segment in the archive used for retrieval is represented in two different forms: lattices in text form generated by the conventional ASR system based on a set of acoustic/language models (at the bottom right corner of Fig. 1), and the one-best lists in acoustic patterns generated by another ASR engine based on acoustic patterns (at the bottom left corner of Fig. 1). Each acoustic pattern is a sequence of phoneme-like acoustic unit, and both of these phoneme-like acoustic units and the word-like acoustic patterns and the acoustic pattern language model are automatically learned from the spoken archive to be retrieved through (lower middle of Fig. 1). When a *text* query is entered, the retrieval engine (middle right of Fig. 1) matches the query terms with the text lattices of the spoken segments to generate the first-pass retrieval result. Because the acoustic patterns are discovered in an unsupervised way, the system does not know which text term an acoustic pattern corresponds to, so the acoustic pattern one-best lists cannot be used in the first-pass retrieval since the query is in text. The first-pass retrieval results are not shown to the user. Instead, top-ranked segments selected as pseudo-relevant segments. The system then extracts the text terms possibly related to the query from these pseudo-relevant segments to generate the expanded query in text form (upper middle of Fig. 1), which gives a new set of retrieval results via the retrieval engine in text (upper left of Fig. 1). Here we have the second version of the expanded query based on acoustic patterns. The acoustic patterns repeatedly occurring in the pseudo-relevant segments, probably corresponding to some query-related terms, are also used to form the second expanded query composed of acoustic patterns to retrieve the spoken segments via the retrieval engine in acoustic patterns, and the results for the two retrieval engines are integrated (upper left of Fig. 1). In this way, **even though some important query-related terms are OOV or incorrectly recognized, the acoustic patterns corresponding to these terms can be included in the expanded query, and the spoken segments containing these acoustic patterns can thus be retrieved.** The results thus obtained are finally shown to the user.

## 2. PROPOSED APPROACH

### 2.1. Off-line Preprocessing

Although any retrieval approach can be used, here we assume the retrieval engine used in Fig 1 is based on the language modeling retrieval approach [21, 22]. The basic idea for this approach is that the query $q$ and the spoken segment $x$ are respectively represented as language models. The segments $x$ are then ranked based on the KL divergence between the language models for the spoken segment $x$ and the query $q$. All language models below are unigram models unless specified otherwise, although the proposed approach is not limited to this case.

At the off-line preprocessing stage, the spoken archive to be retrieved is divided into spoken segments, each with a length of several utterances. Then two different language models are generated for each spoken segment: one based on the lattices generated by the conventional ASR in text, and the other on the one-best lists obtained with another ASR based on automatically discovered acoustic patterns. These language models will be used in the following retrieval and query expansion processes.

#### *2.1.1. Segment Language Models based on Text Terms*

Each segment $x$ is first transcribed into a word lattice and a subword-based lattice (each arc representing a subword unit hypothesis). The expected counts for each term $t$ (a word or a subword unit) are then estimated from the lattices as in (1).

$$E[t|x] = \sum_{u \in \mathcal{L}(x)} N(t,u) P(u|x), \tag{1}$$

where $\mathcal{L}(x)$ is the set of all possible paths in the lattice for $x$, $u$ is a path, $N(t,u)$ the occurrence count of the term $t$ in $u$, and $P(u|x)$ the posterior probability of the term (word or subword unit) sequence $u$ derived from the acoustic and language models.

The term distribution, or (unigram) language model, $\theta_x$ for each spoken segment $x$ is estimated in (2). [1]

$$P(t|\theta_x) = \frac{E[t|x]}{\sum_t E[t|x]}, \tag{2}$$

where $E[t|x]$ is in (1). Then $\theta_x$ is linearly interpolated with a background language model $\theta_b$ based on terms (words or subword units) trained from all spoken segments in the spoken archive $\mathcal{C}$ to form a smoothed model $\tilde{\theta}_x$ [23], where

$$P(t|\theta_b) = \frac{\sum_{x \in \mathcal{C}} E[t|x]}{\sum_t \sum_{x \in \mathcal{C}} E[t|x]} \tag{3}$$

is the probability of observing the term $t$ in the whole archive $\mathcal{C}$.

#### *2.1.2. Segment Language Models based on Acoustic Patterns*

Here we extend the widely studied unsupervised acoustic pattern discovery techniques [24, 25, 26, 27, 28, 29, 30, 31, 32, 33] to find repeated acoustic patterns in the spoken archive. Such techniques have been utilized for enhancing spoken document classification [34, 35, 36, 37], spoken term detection [24, 38, 39, 40], music retrieval [41] and video retrieval [42]; but not yet fully leveraged for semantic retrieval of spoken content. For the approach here each word-like acoustic pattern is a sequence of several phoneme-like acoustic units which are shared by many different word-like acoustic patterns and modeled as HMMs. The transition probabilities between such word-like acoustic patterns are further modeled by an n-gram language model. The two-level phoneme-like acoustic units and word-like acoustic patterns including the alphabet/vocabulary size, the lexicon, and the HMM/n-gram parameters are all automatically learned in an unsupervised way from the spoken archive to be retrieved. These realize a second ASR system completely based on these acoustic patterns of audio signals. This ASR generates a one-best list in acoustic patterns for each spoken segment also used for retrieval. This is the way to try to preserve some of the information in audio signals which may be lost in conventional ASR. Complete presentation for discovering such acoustic patterns is in a companion paper [43]. With the spoken segments transcribed into sequences of acoustic patterns, the counts for the acoustic patterns with label $v$ in each spoken segment $x$, denoted as $C(v,x)$, give a language model $\phi_x$ in (4) based on acoustic patterns [2],

$$P(v|\phi_x) = \frac{C(v,x)}{\sum_v C(v,x)}. \tag{4}$$

[1] The notation $\theta$ indicates a language model based on text terms (words or subword units).

[2] The notation $\phi$ indicates a language model based on acoustic patterns

Then $\phi_x$ is further interpolated by an acoustic pattern background model $\phi_b$ trained from the whole spoken archive exactly as in (3) to form a smoothed version $\bar{\phi}_x$.

### 2.2. First-pass Retrieval

The input text query $q$ can be represented by a term-based[3] language model $\theta_q$,

$$P(t|\theta_q) = \frac{N(t,q)}{|q|}, \tag{5}$$

where $N(t,q)$ is the number of term $t$ in query $q$, and $|q|$ is total number of terms in query $q$. Because the input query is in text form, it can only be matched with the term-based segment models $\bar{\theta}_x$ discussed in Subsection 2.1.1. The language models based on acoustic patterns obtained in Section 2.1.2 cannot be involved here [4]. The relevance score function $S_0(q,x)$ in (6) is used for ranking the spoken segments $x$ for the query $q$ in the first pass.

$$S_0(q,x) = -[(1-w_1)KL(\theta_q^w|\bar{\theta}_x^w) + w_1 KL(\theta_q^s|\bar{\theta}_x^s)], \tag{6}$$

assuming both word-based and subword-based models are used. $S_0(q,x)$ is based on the weighted sum of the KL divergences between word query model $\theta_q^w$ (term $t$ replaced by word $w$ in (5)) and the smoothed word segment model $\bar{\theta}_x^w$ (in parallel with $\bar{\theta}_x$ in Subsection 2.1.1 but for terms being words) and similarly the KL divergence between subword query model $\theta_q^s$ and smoothed subword segment model $\bar{\theta}_x^s$. $w_1$ is the weight. The segment list ranked according to (6) is the first-pass retrieval results used for selecting the pseudo-relevant segments below.

### 2.3. Query Expansion

Although all different approaches for query expansion can be applied, here we adopt and modify the query-regularized mixture model for text information retrieval [13]. 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. This assumption equally applies for words, subwords and acoustic patterns in spoken segments here. Based on this model, query-related words, subwords and acoustic patterns are extracted from the pseudo-relevant spoken segments to form expanded query language models $\theta_{qe}^w$, $\theta_{qe}^s$, and $\phi_{qe}$ respectively for words, subwords and acoustic patterns[5]. The final results shown to the user is ranked according to $S(q,x)$:

$$\begin{aligned} S(q,x) \quad = \quad & -\Big\{(1-w_2)\Big[(1-w_1)KL(\theta_{qe}^w|\bar{\theta}_x^w) + w_1 KL(\theta_{qe}^s|\bar{\theta}_x^s)\Big] \\ & + w_2 KL(\phi_{qe}|\bar{\phi}_x)\Big\}. \end{aligned} \tag{7}$$

In (7), the weight $w_1$ is between scores obtained by word-based and subword-based expanded query language models similar to that in (6), and the weight $w_2$ is between scores based on text terms (words and subwords) and acoustic patterns. $\bar{\theta}_x^w$, $\bar{\theta}_x^s$ and $\bar{\phi}_x$ are respectively the corresponding smoothed segment language models.

[3] word- or subword-based

[4] When input query is spoken, it can be represented by acoustic patterns, but that is out of the scope of this paper.

[5] The subscripts $_{qe}$ indicate the expanded query language models.

#### 2.3.1. Query Expansion for Text Terms

Below is the way to estimate the expanded query model $\theta_{qe}$ for text terms. Suppose the $N$ pseudo-relevant spoken segments are $\{x_1, ..., x_n, ..., x_N\}$. With the assumption that the terms in each pseudo-relevant spoken segment are either query-related or general, the segment language model $\theta_{x_n}$ in (2) should be close to an estimated model $\theta'_{x_n}$ which is the interpolation of the expanded query model $\theta_{qe}$ to be estimated (for query-related words) and the background language model $\theta_b$ in (3) (for general words) with a segment-dependent weight $\alpha_n$.

$$P(t|\theta'_{x_n}) = \alpha_n P(t|\theta_{qe}) + (1-\alpha_n)P(t|\theta_b), \tag{8}$$

where $\alpha_n$ is the segment-dependent interpolation weight for segment $x_n$, which is to be estimated as well. It is therefore reasonable to take the query language model $\theta_{qe}$ minimizing (9) as the expanded query model.

$$F_1(\theta_{qe}, \alpha_1, ..., \alpha_N) = \sum_{n=1}^{N} KL(\theta_{x_n}|\theta'_{x_n}), \tag{9}$$

which means the sum of the KL divergence between each segment model $\theta_{x_n}$ and the corresponding interpolated language model $\theta'_{x_n}$ in (8) for all the $N$ pseudo-relevant segments should be minimized if $\theta_{qe}$ is properly chosen. However, the model $\theta_{qe}$ minimizing (9) may be just for the common content of the pseudo-relevant segments, not necessarily query-related. To better handle this problem, $\theta_{qe}$ is "regularized" by the original query model $\theta_q$ in (5), and we define a function $F_2(\theta_{qe})$ as the prior for $\theta_{qe}$ based on $\theta_q$,

$$F_2(\theta_{qe}) = KL(\theta_q|\theta_{qe}). \tag{10}$$

$F_2(\theta_{qe})$ will be smaller for model $\theta_{qe}$ closer to $\theta_q$. The expanded query model $\theta_{qe}$ and the weight $\alpha_n$ are actually estimated by minimizing the following objective function:

$$F(\theta_{qe}, \alpha_1, ..., \alpha_N) = F_1(\theta_{qe}, \alpha_1, ..., \alpha_N) + \lambda F_2(\theta_{qe}), \tag{11}$$

where $\lambda$ is a parameter controlling the influence of the function $F_2(\theta_{qe})$. The model $\theta_{qe}$ estimated via minimizing (11) would not be totally drifted away by the pseudo-relevant segments because the function $F_2(\theta_{qe})$ prefers the expanded query model $\theta_{qe}$ to be similar to the original query model $\theta_q$. The above is equally applied to words or subword units, and gives the expanded query models $\theta_{qe}^w$ and $\theta_{qe}^s$ in (7).

#### 2.3.2. Query Expansion for Acoustic Patterns

Equation (11) above cannot be directly used to estimate the expanded query model based on acoustic patterns $\phi_{qe}$ because $F_2(.)$ in (10) is undefined, or the input text query can not be represented in acoustic patterns. We can simply find $\phi_{qe}$ and $\alpha'_n$ minimizing (12),

$$F'_1(\phi_{qe}, \alpha'_1, ..., \alpha'_N) = \sum_{n=1}^{N} KL(\phi_{x_n}|\phi'_{x_n}), \tag{12}$$

where

$$P(v|\phi'_{x_n}) = \alpha'_n P(v|\phi_{qe}) + (1-\alpha'_n)P(v|\phi_b), \tag{13}$$

which is the interpolation of $\phi_{qe}$ and the background model $\phi_b$ in Section 2.1.2 with weight $\alpha'_n$, and (12) and (13) are very similar to (9) and (8). However, without the query regularization, the above estimation may be risky, or $\phi_{qe}$ may be drifted away. Here we assume

the acoustic patterns approximately correspond to the text terms, so for a spoken segment $x_n$, the ratio $\alpha'_n$ in (13) for acoustic patterns and the ratio $\alpha_n$ in (8) for text terms should be close. Hence, when minimizing (12), it may be reasonable to set $\alpha'_n = \alpha_n$ in (13) instead of trying to estimate them, and therefore only $\phi_{qe}$ is estimated. This yields more robust expanded queries based on acoustic patterns than without query regularization.

## 3. EXPERIMENTS

### 3.1. Experimental Setup

The spoken archive to be retrieved in the experiments consisted of 4 hours of Mandarin broadcast news stories collected daily from local radio stations in Taiwan in 2001. We manually segmented these stories into 5034 spoken segments, each with one to three utterances. A trigram language model estimated from a 40M news corpus collected in 1999 and a lexicon of 62K words was used for recognition. The acoustic models used included 151 intra-syllable right-context-dependent Initial-Final models for Mandarin syllables, trained using 8 hours of broadcast news stories collected in 2000. The beam width for recognition was 60, and the one-best recognition character accuracy for the spoken archive was 75.27%. After each spoken segment was transcribed into a word lattice, we further transformed each Chinese word arc in the lattice into a sequence of concatenated corresponding Chinese character arcs to form character lattices, or characters are taken as subword units here. 29 single word in-vocabulary queries were manually selected for the retrieval experiments. The corresponding semantically relevant spoken segments were manually selected, which did not necessarily contain the queries. Mean average precision (MAP) was used as the performance measure [44].

### 3.2. Experimental Results

Fig. 2 (a) shows the MAP yielded by (7), integrating the results of expanded queries based on text terms ($\theta^w_{qe}$ for words and $\theta^s_{qe}$ for characters) and acoustic patterns ($\phi_{qe}$) with different numbers of pseudo-relevant segments ($N = 5, 10, 15, 20, 25$). The red line in the figure is the MAP of the first-pass retrieval results without query expansion. The horizontal scale is for different values of the interpolation weight $w_2$ for the acoustic patterns in (7) with $w_2 = 0$ for the results without acoustic patterns. $w_1$ in (6) and (7) was fixed to 0.95 [6], and $\lambda$ in (11) was 800. First of all, comparing the MAP of the first-pass results and the query expansion based on text terms only (red line vs $w_2 = 0$), we found that query expansion based on text terms offered some improvements over the baseline even though the recognition errors and OOV problems probably limited its performance. With the help of the acoustic patterns, extra improvements over the query expansion based on text terms were always achieved ($w_2 > 0$ vs $w_2 = 0$) as long as $w_2$ was smaller than 0.5 regardless of the number of pseudo-relevant segments ($N$). This verified that the acoustic patterns directly charactering speech signals are really helpful for query expansion in spoken content retrieval. Also, as $N$ was raised, the MAP first increased and then decreased. Larger $N$ implies more segments considered, and more training data used in (11) and (12). However, when $N$ was too large, more irrelevant segments were inevitably included in the pseudo-relevant segment set and disturbed the estimation thereby. The best result was for $N = 10$ and $w_2 = 0.40$, which means query expansion based on text terms were more precise than that based on acoustic patterns. $N$ was thus set to 10 in the following experiments.

[6] This setting yielded the best results for query expansion based on text terms only, or $w_2 = 0$.

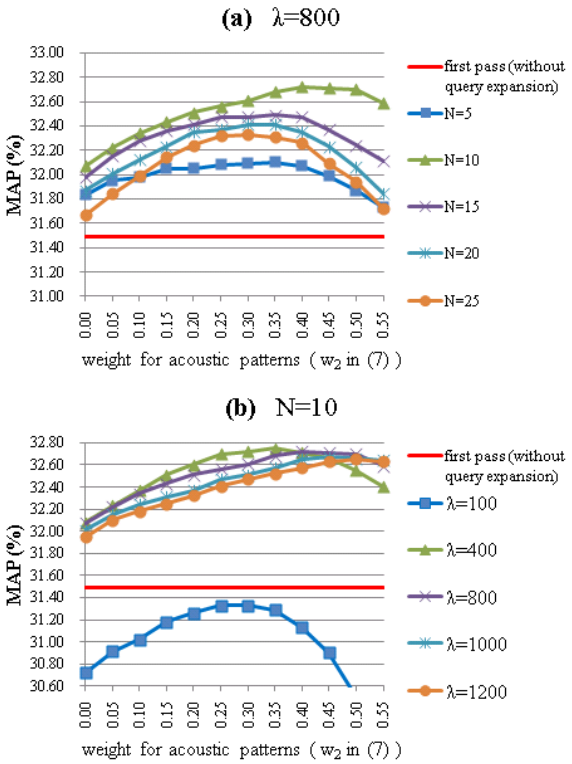


**Fig. 2**. *MAP yielded by integrating query expansion based on text terms (words plus characters) and acoustic patterns. (a) for different numbers of pseudo-relevant segments ($N = 5, 10, 15, 20, 25$) with $\lambda = 800$ and (b) for $N = 10$ and different values of $\lambda$ in (11). The red lines in the figures are the MAP of the first-pass retrieval results. The horizontal scale is the interpolation weight $w_2$ in (7), and $w_2 = 0$ is the case without acoustic patterns.*

Fig. 2 (b) is exactly the same as Fig. 2 (a), except that $N = 10$ and different values of $\lambda$ in (11) were tested. We found that small $\lambda$ ($\lambda = 100$) yielded very poor results even worse than the first pass obviously because the query model was drifted too much by the pseudo-relevant segments. However, when $\lambda$ was large enough ($\lambda \geq 400$), improvements over the first pass and $w_2 = 0$ were always observed. This verified the importance of the regularization term in (11), while the acoustic patterns always improved the performance of query expansion ($w_2 > 0$ vs $w_2 = 0$) when $\lambda$ was large enough.

## 4. CONCLUSION

In this paper, we used acoustic patterns discovered from the spoken content to enhance the query expansion techniques originally developed for text information retrieval. The usefulness of the proposed approach were verified on a Mandarin broadcast news corpus.

## 5. REFERENCES

[1] Ciprian Chelba, Timothy J. Hazen, and Murat Saralar, "Retrieval and browsing of spoken content," in *IEEE Signal Processing Magazine 25(3), pp. 39-49*, 2008.

[2] Lin-Shan Lee and Chen B., "Spoken document understanding and organization," *Signal Processing Magazine, IEEE*, vol. 22, pp. 42 – 60, 2005.

[3] *http://www.itl.nist.gov/iad/mig/tests/std/2006/index.html*.

[4] 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.

[5] Tsung-Wei Tu, Hung-Yi Lee, Yu-Yu Chou, and Lin-Shan Lee, "Semantic query expansion and context-based discriminative term modeling for spoken document retrieval," in *ICASSP*, 2012.

[6] Hung lin Chang, Yi cheng Pan, and Lin-Shan Lee, "Latent semantic retrieval of spoken documents over position specific posterior lattices," in *SLT*, 2008.

[7] B. Chen, K.-Y. Chen, P.-N. Chen, and Y.-W. Chen, "Spoken document retrieval with unsupervised query modeling techniques," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2602 –2612, nov. 2012.

[8] Xinhui Hu, Ryosuke Isotani, Hisashi Kawai, and Satoshi Nakamura, "Cluster-based language model for spoken document retrieval using NMF-based document clustering," in *Interspeech*, 2010.

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

[10] Ryo Masumura, Seongjun Hahm, and Akinori Ito, "Language model expansion using webdata for spoken document retrieval," in *Interspeech*, 2011.

[11] Hiromitsu Nishizaki, Kiyotaka Sugimotoy, and Yoshihiro Sekiguchi, "Web page collection using automatic document segmentation for spoken document retrieval," in *APSIPA*, 2011.

[12] S. Tsuge, H. Ohashi, N. Kitaoka, K. Takeda, and K. Kita, "Spoken document retrieval method combining query expansion with continuous syllable recognition for NTCIR-SpokenDoc," in *Proceedings of the Ninth NTCIR Workshop Meeting*, 2011.

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

[14] Victor Lavrenko and W. Bruce Croft, "Relevance-based language models," in *SIGIR*, 2001.

[15] Yuanhua Lv and ChengXiang Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, CIKM '09.

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

[17] Atta Norouzian, Aren Jansen, Richard Rose, and Samuel Thomas, "Exploiting discriminative point process models for spoken term detection," in *Interspeech*, 2012.

[18] Hung-Yi Lee, Po-Wei Chou, and Lin-Shan Lee, "Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity," in *Interspeech*, 2012.

[19] Hung-Yi Lee, Yun-Nung Chen, and Lin-Shan Lee, "Improved speech summarization and spoken term detection with graphical analysis of utterance similarities," in *APSIPA*, 2011.

[20] Hung-Yi Lee, Chia-Ping Chen, and Lin-Shan Lee, "Integrating recognition and retrieval with relevance feedback for spoken term detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 2095 –2110, 2012.

[21] ChengXiang Zhai, "Statistical language models for information retrieval a critical review," *Found. Trends Inf. Retr.*, vol. 2, no. 3, pp. 137–213, Mar. 2008.

[22] Tee Kiah Chia, Khe Chai Sim, Haizhou Li, and Hwee Tou Ng, "Statistical lattice-based spoken document retrieval," *ACM Trans. Inf. Syst.*, vol. 28, pp. 2:1–2:30, 2010.

[23] Chengxiang Zhai and John Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *SIGIR*, 2001.

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

[25] Aren Jansen and Kenneth Church, "Towards unsupervised training of speaker independent acoustic models," in *Interspeech*, 2011.

[26] Aren Jansen, Kenneth Church, and Hynek Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

[27] A.S. Park and J.R. Glass, "Unsupervised pattern discovery in speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 186 –197, jan. 2008.

[28] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorization," *Signal Processing Letters, IEEE*, vol. 15, pp. 131 –134, 2008.

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

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

[31] J. Driesen and H. Van hamme, "Fast word acquisition in an NMF-based learning framework," in *ICASSP*, 2012.

[32] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorization," *Signal Processing Letters, IEEE*, vol. 15, pp. 131 –134, 2008.

[33] Yaodong Zhang and J.R. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *ICASSP*, 2010.

[34] 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.

[35] Timothy J. Hazen, Man-Hung Siu, Herbert Gish, Steve Lowe, and Arthur Chan, "Topic modeling for spoken documents using only phonetic information," in *ASRU*, 2011.

[36] Herbert Gish, Man hung Siu, and Arthur Chan amd William Belfield, "Unsupervised training of an HMM-based speech recognizer for topic classification," in *Interspeech*, 2009.

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

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

[39] Chun-An Chan and Lin-Shan Lee, "Unsupervised hidden Markov modeling of spoken queries for spoken term detection without speech recognition," in *Interspeech*, 2011.

[40] 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.

[41] Matthew Riley, Eric Heinen, and Joydeep Ghosh, "A text retrieval approach to content-based audio retrieval," in *ISMIR*, 2008.

[42] 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.

[43] 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.

[44] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees, "The TREC spoken document retrieval track: A success story," in *Text Retrieval Conference (TREC) 8*, 2000.