Improved Semantic Retrieval of Spoken Content by Document/Query Expansion with Random Walk over Acoustic Similarity Graphs

Hung-yi Lee and Lin-shan Lee, *Fellow, IEEE*

***Abstract*—In a text context, document/query expansion has proven very useful in retrieving objects semantically related to the query. However, when applying text-based techniques on spoken content, the inevitable recognition errors seriously degrade performance even when the retrieval process is performed over lattices. We propose the estimation of more accurate term distributions (or unigram language models) for the spoken documents by acoustic similarity graphs. In this approach, a graph is constructed for each term describing the acoustic similarity among all signal regions hypothesized to be the considered term. Score propagation based on a random walk over the graph offers more reliable scores of the term hypotheses, which in turn yield more accurate term distributions (or unigram language models). This approach was applied with the language modeling retrieval approach, including using document expansion based on latent topic analysis and query expansion with a query-regularized mixture model. We extend these approaches from words to subword n-grams, and the query expansion from document-level to utterance-level and from term-based to topic-based. Experiments performed on Mandarin broadcast news showed improved performance under almost all tested conditions.**

***Index Terms*—Spoken Content Retrieval, Random Walk, Latent Semantic Analysis, Query Expansion, Document Expansion**

I. Introduction

In recent years, the demand for information in our daily lives has clearly gone beyond traditional text information [1]. With the ever-increasing bandwidth of the Internet and rapidly falling storage costs, multimedia data such as shared videos, broadcast programs, lectures, meeting records, and many other video/audio materials are now the most widely accessed network content. However, compared to text, multimedia/audio content is difficult to retrieve and browse, while the speech information included in such content very often indicates its subject or topic. This underscores the need for efficient technologies for retrieving spoken content, which will provide users with easy access to the huge quantities of multimedia/audio resources over the Internet.

Substantial effort has been made in spoken content retrieval in recent years, and many successful techniques have been developed [2], [3]. Lattice-based approaches that take into account multiple recognition hypotheses [4], [5] have been used to take mitigate the relatively low accuracy in one-best transcriptions. In some example approaches, lattices have been compressed into more compact structures such as position-specific posterior lattices (PSPL) [6], [7] and confusion networks (CN) [7], [8] to facilitate indexing and save on memory space. Weighted finite state transducer (WFST) algorithms have provided another effective way to index and retrieve lattices [9], [10]. Out-of-vocabulary (OOV) queries represent another important problem because typically many queries contain OOV terms [11]. The most fundamental approach for handling the OOV problem is to represent both the queries and the spoken content by properly chosen subword units and then to try to match them at the subword level [12]–[21]. Word-based and subword-based indexing can be further integrated to yield better performance [13], [22], [23]. Many successful applications have been demonstrated with good examples including those browsing over broadcast news [24], [25], course lectures [26], [27], historical spoken archives [28], podcasts [29], and YouTube videos [30]. However, most works in spoken content retrieval remain focused on spoken term detection (STD), for which the goal is simply returning spoken segments that include the query terms. This is insufficient because users naturally expect the technologies to return all the objects they are looking for, regardless of whether the query terms are included or not. For example, when the user enters the query “airplane”, a system that returns only spoken documents including the query term “airplane” but not those containing the related term “aircraft” may not meet the user’s information needs. This consideration has led to extensive recent work on the semantic retrieval of spoken content [31]–[39].

There are in general two stages for semantic retrieval of spoken content. In the first stage, the audio content is recognized and transformed into transcriptions or lattices based on a set of acoustic models and language models. In the second stage, after the user enters a query, the retrieval engine searches through the recognition output and returns a list of relevant spoken documents to the user. Taking the one-best transcriptions as the text for the content, any technique developed for text information retrieval, such as language modeling retrieval and query/document expansion [40]–[45], can be directly applied to the semantic retrieval of spoken content. Language modeling retrieval has been shown very effective for information retrieval not only for text, but for spoken content as well [31], [32], [34], [46]. Although it does not directly address the problem of relevant documents that do not contain the query terms, it provides a reasonable framework on top of which other advanced techniques can be applied in addition. With document expansion, to handle the problem of term usage mismatch between the queries and the documents semantically related to the queries, each text or spoken document is expanded by those terms related to its content based on latent topic analysis [31], [34], [35]. Query expansion offers another effective way to retrieve semantically related documents in text or spoken form; this enriches the representation of short queries with some related terms. Automatic query expansion techniques widely studied in text information retrieval such as the relevance model and the query-regularized mixture model have been successfully applied to spoken content retrieval as well [31], [32], [34]. External information from the web is also helpful for the expansion of both spoken documents and queries [36]–[39].

Although the above techniques seem promising, because these techniques for spoken content retrieval were originally developed for text without errors, the inevitable recognition errors and resulting uncertainty may seriously degrade performance. One way to handle the problem of recognition errors is to consider multiple recognition hypotheses of spoken documents using lattice structures. However, when the acoustic models and language models used in the recognition are highly mismatched to the target spoken archive, even though the correct word hypotheses may be included in the lattices, incorrect noisy hypotheses may make it very difficult to extract the desired information from the spoken documents.

For spoken term detection, it has been found that graph-based re-ranking using acoustic feature similarity between query hypotheses is very helpful [47]–[49]. The basic assumption for this approach is that acoustic feature sequences representing different occurrences of the same term may be similar in some aspects, and consequently that very different feature sequences are probably different terms. Therefore, for each given user query, all signal regions hypothesized to be the query term with confidence scores are used to construct a graph in which each node represents a signal region hypothesized to be the query term, and the edge weights are the similarities between the acoustic feature sequences for the two corresponding nodes. Based on the above assumption, nodes strongly connected to more nodes with higher confidence scores on the graph should have higher confidence scores; confidence scores for the nodes thus propagate over the graph, yielding better detection results.

In this paper, we use a similar concept of graphs of acoustic similarity to estimate more accurately the language models for the spoken documents for better semantic retrieval. We first verify that the proposed approach improves the performance of the standard language modeling retrieval approach, because better language models for the spoken documents enhanced with the graphs of acoustic similarity lead directly to better retrieval. This language modeling retrieval approach can be otherwise improved by document expansion based on topic analysis and query expansion based on the query-regularized mixture model, and we show that under such conditions the proposed approach offered additional improvements. The document/query expansion and the proposed approach are complementary to each other because document/query expansion focuses on retrieving the relevant documents without the query terms, while the spoken documents are better represented by the proposed approach. In addition, the approach can be equally applied to different granularities of terms including words, subword units, or several consecutive words or subwords, and information from different granularities of terms can be fused to improve retrieval performance. Furthermore, we also extend the query-regularized mixture model from the document level [43] to the utterance level and incorporate latent topic information with query expansion.

The remainder of this paper is structured as follows. We present the language modeling approach for spoken content retrieval and the proposed graph-based enhancement approach in Section II. Document and query expansion approaches for spoken content are then described in Sections III and IV respectively. Experiments are reported in Sections V and VI, and in Section VII are the concluding remarks.

II. Language Modeling Retrieval Approach for Spoken Content

Here we start with the language modeling retrieval approach using one-best transcriptions in Section II-A, which is exactly the same as the conventional language modeling retrieval approach for text information retrieval. Then we explain how it is extended to spoken content with lattices in Section II-B, and present the proposed graph-based enhancement approach in Section II-C.

A. Conventional Language Modeling Retrieval Approach

The language modeling retrieval approach has been shown to be very effective for both text and speech information retrieval [31], [32], [34], [46]. The conventional language modeling approach for text can be directly applied on spoken content as long as the spoken content is transcribed into one-best transcriptions. The basic idea for this approach is that the query Q and document d are respectively represented as unigram language models θ_Q and θ_d, or term distributions $P(t|\theta_Q)$ and $P(t|\theta_d)$, where t is a term[1]. The relevance score function $S(Q, d)$ used to rank the documents d with respect to the given query Q is the inverse of the KL divergence between θ_Q and θ_d:

$$S(Q, d) = -KL(\theta_Q||\theta_d). \tag{1}$$

That is, documents whose unigram language models are similar to the query's unigram language model are more likely to be relevant. In this way, the problem of retrieval is reduced to the estimation of the unigram language models for the queries and documents. We here assume that the term t can be a sequence of n consecutive words or subword units (referred to as word or subword n-gram, or a word or a subword unit if $n = 1$). In other words, here θ_Q and θ_d are the distributions of such word or subword n-grams. Therefore, although θ_Q and θ_d are just unigram language models in the following discussion, the context information can be naturally considered

[1] In the following discussion, we assume that both θ_Q and θ_d are unigram language models, although the language modeling retrieval approach is not limited to this case. Although there is ongoing work to extend the language model from just unigrams to also include n-grams and grammars, this has yielded only mild gains [50].

with longer n-grams, and the problem of OOV queries can be properly handled with subword units. The results based on word and subword n-grams can be properly integrated to yield better results.

Here we focus only on queries in text form. Although it is certainly possible to extend the proposed approaches to spoken queries if they are transcribed into text symbols, this is out of the scope of this paper. Usually the unigram language model θ_Q for the query Q is estimated based on the terms in Q as

$$P(t|\theta_Q) = \frac{N(t,Q)}{\sum_t N(t,Q)}, \tag{2}$$

where $P(t|\theta_Q)$ is the probability of generating the term t from the model θ_Q, and $N(t,Q)$ the occurrence counts of the term t in Q. The denominator of (2) normalizes $N(t,Q)$ into probabilities $P(t|\theta_Q)$.

Even though the documents considered here are spoken, when they are transcribed into one-best transcriptions, the estimation of a document's unigram language model is exactly the same as that for text. A document's unigram language model θ_d^{1b} is first estimated based on the terms in the one-best transcriptions of spoken document d in (3) below. The superscript ^{1b} indicates that the unigram language models are directly derived from the one-best transcriptions. Then θ_d^{1b} is interpolated with a background model θ_b^{1b} in (4) trained from the one-best transcriptions of all the spoken documents in the archive to be retrieved from to form a smoothed document model $\bar{\theta}_d^{1b}$ in (5).

$$P(t|\theta_d^{1b}) = \frac{N(t,d)}{\sum_t N(t,d)}, \tag{3}$$

where $N(t,d)$ is total counts for term t in the one-best transcriptions of d, and the denominator is the summation over all terms t.

$$P(t|\theta_b^{1b}) = \frac{N(t,\mathcal{C})}{\sum_t N(t,\mathcal{C})}, \tag{4}$$

where $\mathcal{C}$ represents the whole spoken document collection to be retrieved from, and $N(t,\mathcal{C})$ is the total count for term t in the one-best transcriptions of $\mathcal{C}$.

$$P(t|\bar{\theta}_d^{1b}) = \lambda_d P(t|\theta_d^{1b}) + (1-\lambda_d)P(t|\theta_b^{1b}), \tag{5}$$

where $\lambda_d = \frac{L_d}{L_d+\kappa}$ is a document-dependent interpolation weight, and κ is a parameter to be tuned. L_d is the document length, which if the term t being considered is a word n-gram is the total number of words in d, or if term t is a subword n-gram[2] is the total number of subwords in d. With the document-dependent weight λ_d, the background model has more influence on shorter documents. The same smoothing strategies have been widely applied in text information retrieval: like the well-known inverse document frequency [51], it has been shown that such smoothing strategies implicitly weight rare but informative terms. The smoothed model $\bar{\theta}_d^{1b}$ in (5) is used for θ_d in (1) for ranking. Due to the inevitable high rate of errors in one-best transcriptions, $\bar{\theta}_d^{1b}$ thus estimated may vary widely from the true word distribution of the spoken document.

B. Language Modeling Retrieval Approach for Spoken Content based on Lattices

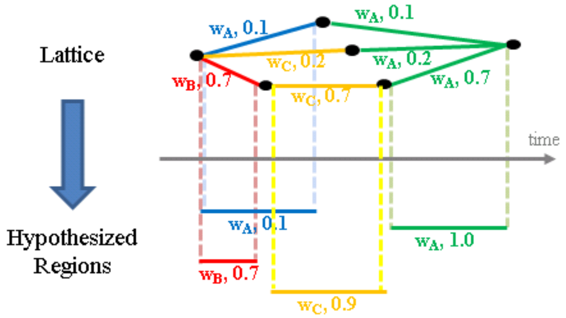


Fig. 1: *Hypothesized regions of terms in a lattice. The upper part is a word lattice of a spoken segment, and the word hypotheses and posterior probabilities of the arcs are shown beside each arc. The lower part is the hypothesized regions of terms obtained from the lattice; the confidence scores are shown beside each hypothesized region.*

For spoken content, lattices are better representations than one-best transcriptions, so each document in the speech archive to be retrieved through is first divided into spoken segments, each of which is decoded into lattice form. The lattices can be either word- or subword-based, or the arcs in the lattices can be either word or subword hypotheses. Since the paths embedded in the lattices have different acoustic likelihoods and language model scores, the occurrence frequencies of the terms in the spoken segments cannot be directly counted from the lattices. Here we first define the hypothesized regions for the terms from the lattices of the spoken segments. The document models are then estimated using these hypothesized regions.

Fig. 1 shows how we define the hypothesized regions for word 1-grams, or single words, from a word lattice. This can be directly extended to longer n-grams and subword-based lattices. The upper part of Fig. 1 is a word lattice of a spoken segment: the word hypotheses and posterior probabilities for the arcs are shown beside each arc. The lower part of Fig. 1 is the hypothesized regions for the terms obtained from the lattice. First, each arc in the word lattice in the upper half of Fig. 1 corresponds to a possible term occurrence. Then arcs corresponding to the same term with similar time spans are clustered into a single group to define a single hypothesized region of the term. Each hypothesized region has a representative time span and a confidence score describing the posterior probability that the term occurs in this time span (also shown beside each hypothesized region in the lower part of Fig. 1). The detailed procedure for this is presented below. In general, when word/subword n-grams are taken as terms, each arc sequence with n consecutive arcs in the word/subword-based lattices corresponds to a possible occurrence of a term; everything else is handled similarly.

[2] For a document d with L_d words, there are exactly $L_d-(n-1)$ different word n-grams. Therefore, $L_d-(n-1)$ is a more precise definition of the document length. However, because L_d considered here is relatively large but n considered here is usually small, we simply use L_d as the document length.

Below we present the algorithm to obtain from a lattice the hypothesized regions for a term under consideration. For all the terms t in the lattice of a spoken segment x, we conduct the following steps to find the hypothesized regions of t:

(i) Collect all the arc sequences a corresponding to t in the lattice of segment x to form a set $\mathcal{A}_t$. Each arc sequence has a posterior probability $P(a|x)$ derived from the acoustic likelihoods and language model scores from the lattice for x:

$$P(a|x) = \sum_{u \in \mathcal{L}(x), a \in u} P(u|x), \qquad (6)$$

where u is a path in the lattice, and $\mathcal{L}(x)$ the set of all possible paths in the lattice for x. $P(u|x)$ is the posterior probability of path u,

$$P(u|x) = \frac{P(x|u)P(u)}{\sum_{u \in \mathcal{L}(x)} P(x|u)P(u)}, \qquad (7)$$

where $P(x|u)$ is the likelihood for the observation sequence of x given the path u based on the acoustic model set, and $P(u)$ the prior probability of u from the language model. Therefore, $P(a|x)$ is the sum of the posterior probabilities of all paths in the lattice of x which include the arc sequence a.

(ii) Find the arc sequence a^* in $\mathcal{A}_t$ with the largest posterior probability,

$$a^* = arg \max_a P(a|x). \qquad (8)$$

(iii) Find all arc sequences $\hat{a}$ in $\mathcal{A}_t$ whose time spans include the center of the time span of a^* to form a subset $\hat{\mathcal{A}}_t$ ($a^* \in \hat{\mathcal{A}}_t$ naturally).

(iv) Then the hypothesized region r_t for the t is defined as the time span of a^* with a confidence score (or the confidence that t actually occurs in this time span) $C(r_t)$:

$$C(r_t) = \sum_{\hat{a} \in \hat{\mathcal{A}}_t} P(\hat{a}|x). \qquad (9)$$

(v) Remove all the arc sequences in $\hat{\mathcal{A}}_t$ from $\mathcal{A}_t$. Go to step (ii) until no elements are left in $\mathcal{A}_t$. There can be more than one hypothesis region for a given term t in a lattice.

The document model θ_d^{lat} for a document d composed of N spoken segments, $d = \{x_1, \cdots, x_n, \cdots, x_N\}$, is then estimated in (10). The superscript ^{lat} indicates that the models are directly derived from the lattices.

$$P(t|\theta_d^{lat}) = \frac{\sum_{n=1}^{N} \sum_{r_t \in x_n} C(r_t)}{\sum_t \sum_{n=1}^{N} \sum_{r_t \in x_n} C(r_t)}, \qquad (10)$$

where the expression $r_t \in x_n$ means the time span of r_t, a hypothesized region of term t, is within the time span of the spoken segment x_n. The numerator of (10) is the sum of the confidence scores $C(r_t)$ for all hypothesized regions of t in all the spoken segments in d, and the denominator normalizes it into a probability. Note that because $C(r_t)$ is the sum of the posterior probabilities for a set of arc sequences corresponding to hypothesized region r_t for term t as in (9), and the term $\sum_{r_t \in x_n} C(r_t)$ in the numerator of (10) is the summation over all r_t in the lattice of a spoken segment x_n, $\sum_{r_t \in x_n} C(r_t)$ is exactly the expected term frequency of t based on the lattice of the spoken segment x_n [52]. We separate the arc sequences corresponding to t into several different hypothesized regions r_t only for the purpose of the proposed graph-based approach described in the next subsection.

The document model θ_d^{lat} in (10) is then interpolated with a background model θ_b^{lat} in (11) obtained in a similar way as in (4) but from the lattices of all the spoken segments in the document collection $\mathcal{C}$ to form a smoothed model $\bar{\theta}_d^{lat}$ in (12).

$$P(t|\theta_b^{lat}) = \frac{\sum_{r_t \in \mathcal{C}} C(r_t)}{\sum_t \sum_{r_t \in \mathcal{C}} C(r_t)}, \qquad (11)$$

where the expression $r_t \in \mathcal{C}$ means $r_t \in x_n$, $x_n \in d$, and $d \in \mathcal{C}$, so the numerator considers all hypothesized regions of term t in the whole collection C, and the denominator further sums over all the terms.

$$P(t|\bar{\theta}_d^{lat}) = \lambda_d' P(t|\theta_d^{lat}) + (1 - \lambda_d') P(t|\theta_b^{lat}), \qquad (12)$$

where $\lambda_d' = \frac{L_d'}{L_d' + \kappa}$ is another document-dependent interpolation weight. κ is the parameter used in λ_d in (5). Since the lattices are considered here, the document length L_d' is the expected length of the document d estimated from the lattices of d.

$$L_d' = \sum_{n=1}^{N} L_{x_n}, \qquad (13)$$

where

$$L_{x_n} = \sum_{u \in \mathcal{L}(x_n)} |u| P(u|x_n), \qquad (14)$$

where $|u|$ is the number of arcs (or number of words for word lattices and number of subword units for subword-based lattices in the path u). This smoothed model $\bar{\theta}_d^{lat}$ in (12) is finally used for θ_d in (1) for ranking.

C. Graph-enhanced Document Model based on Acoustic Similarity

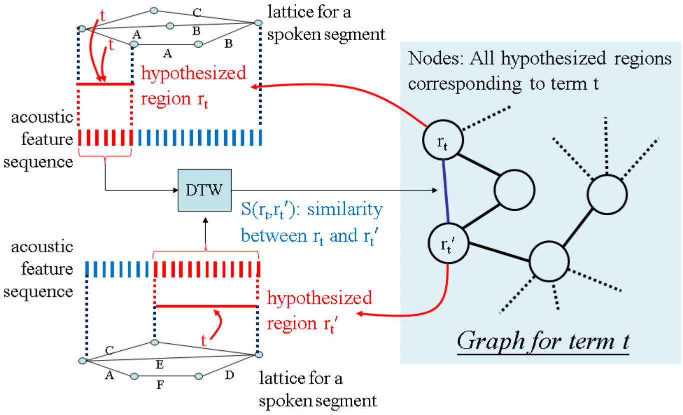

t
A
C
B
B
A
t
lattice for a spoken segment
hypothesized region r_t
acoustic feature sequence
Nodes: All hypothesized regions corresponding to term t
r_t
DTW
$S(r_t, r_t')$: similarity between r_t and r_t'
r_t'
acoustic feature sequence
hypothesized region r_t'
t
C
E
A
F
D
lattice for a spoken segment
Graph for term t

Fig. 2: *The graph constructed for all hypothesized regions corresponding to term t in the spoken document collection $\mathcal{C}$. Each node in the graph represents a hypothesized region corresponding to t, and the edge weights represent the acoustic similarities between the nodes.*

Although the document models derived above in (12) from the lattices may be better than the ones from the one-best transcriptions, they still inevitably suffer from the recognition errors and noisy hypotheses in the lattices. Note that when transcribing speech signals into lattices, much information in the signals is already lost and not recoverable. A possible approach to remedy this problem is to incorporate additional acoustic feature information into the language models derived above to reflect signal-level information. This may be achieved based on an assumption that the acoustic feature segments of different occurrences for the same term may appear somewhat similar because they all include very similar phoneme segments. Here we make this assumption in an attempt to improve the document models derived above in (12). Similar approaches have been applied on spoken term detection (STD) [47]–[49], [53]–[55].

In this approach, for each term t appearing in the lattices for the spoken document collection $\mathcal{C}$, we first construct a graph for t using all the hypothesized regions r_t (the hypothesized regions of t) in $\mathcal{C}$, in which each node represents a hypothesized region r_t in $\mathcal{C}$. Such a graph is shown in Fig. 2. Dynamic time warping (DTW) is performed between the acoustic feature sequences corresponding to all the hypothesized region pairs r_t and r'_t on the graph. This yields $d(r_t, r'_t)$, the DTW distance between hypothesized regions r_t and r'_t. The similarity between r_t and r'_t is then defined as

$$S(r_t, r'_t) = 1 - \frac{d(r_t, r'_t) - d_{min}}{d_{max} - d_{min}}, \quad (15)$$

where d_{max} and d_{min} are the largest and smallest values of $d(r_t, r'_t)$ for all node pairs on the graph. Equation (15) linearly normalizes the DTW distance and transforms it into a similarity score between 0 and 1. Only those node pairs for which $S(r_t, r'_t)$ exceeds a threshold are then connected with an edge with weight $S(r_t, r'_t)$.

Note that in the above graph each node r_t has an original confidence score $C(r_t)$ derived previously from the lattices in (9). Now with this graph, the above assumption that the acoustic features for different occurrences of the same term may be similar implies that those nodes on the graph (or hypothesized regions for term t) connected to many other nodes with large edge weights on the graph (or other hypothesized regions for t in the collection $\mathcal{C}$) with higher (or lower) confidence scores should have higher (or lower) confidence scores. In other words, the confidence scores $C(r_t)$ on the graph may "propagate" over the graph through connected edges, or the confidence scores of the hypothesized regions may be "smoothed" over the graph to produce a new set of better confidence scores $C^g(r_t)$. The superscript ^{g} indicates *graph-enhanced*.

The above score propagation can be formulated as trying to find a new set of confidence scores $C^g(r_t)$ satisfying (16) for all r_t on the graph.

$$C^g(r_t) = (1-\alpha)C(r_t) + \alpha \sum_{r'_t \in \mathcal{E}(r_t)} C^g(r'_t)\hat{S}(r'_t, r_t), \quad (16)$$

where $C^g(r'_t)$ is the graph-enhanced confidence score, α an interpolation weight between 0 and 1, $\mathcal{E}(r_t)$ the set of all hypothesized regions r'_t connected to r_t, and $\hat{S}(r'_t, r_t)$ the edge weight normalized over all edges connected to node r'_t on the graph:

$$\hat{S}(r'_t, r_t) = \frac{S(r'_t, r_t)}{\sum_{r''_t \in \mathcal{E}(r'_t)} S(r'_t, r''_t)}. \quad (17)$$

The second term in (16) describes the score propagation on the graph: the node of r_t absorbs the scores from all nodes r'_t connected to r_t ($r'_t \in \mathcal{E}(r_t)$) but weighted by $\hat{S}(r'_t, r_t)$, while the score of r'_t is distributed to all nodes r''_t connected to r'_t ($r''_t \in \mathcal{E}(r'_t)$) as in (17). Therefore, in (16) the graph-enhanced confidence score $C^g(r_t)$ depends on two factors interpolated by α: the original confidence score $C(r_t)$ in (9) (the first term on the right hand side of (16)) and the score propagation over the graph (the second term on the right hand side). Thus $C^g(r_t)$ is larger if $C(r_t)$ is larger, or if r_t is strongly connected to other hypothesized regions r'_t with larger $C^g(r'_t)$ on the graph.

Equation (16) is actually a random walk problem on the graph. Random walk theory guarantees that the score propagation over the graph converges and a set of unique solutions of $C^g(r_t)$ can be found by the power method [56]. In this method, each node r_t is first given an initial value $C^0(r_t)$[3]. Then at each iteration l, $C^{l-1}(r_t)$ obtained in the last iteration is updated to $C^l(r_t)$ as

$$C^l(r_t) = (1-\alpha)C(r_t) + \alpha \sum_{r'_t \in \mathcal{E}(r_t)} C^{l-1}(r'_t)\hat{S}(r'_t, r_t). \quad (18)$$

Equation (18) is parallel to (16), except that here $C^l(r_t)$ rather than $C^g(r_t)$ is at the left hand side of the equation, and $C^{l-1}(r'_t)$ replaces $C^g(r'_t)$ at the right hand side. Whenever the results converge, that is, $C^{l-1}(r_t)$ and $C^l(r_t)$ are sufficiently close, $C^l(r_t)$ can be taken as the scores $C^g(r_t)$ which satisfy (16). The above process of graph construction and score propagation by random walk is performed off-line for all terms t appearing in the lattices for all spoken documents in the collection $\mathcal{C}$.

The graph-enhanced language model θ^g_d for document d is then obtained as

$$P(t|\theta^g_d) = \frac{\sum_{n=1}^{N} \sum_{r_t \in x_n} C^g(r_t)}{\sum_t \sum_{n=1}^{N} \sum_{r_t \in x_n} C^g(r_t)}, \quad (19)$$

which is exactly parallel to (10) except that $C(r_t)$ is replaced by $C^g(r_t)$. Note that if α is set to 0, that is, the score propagation over the graph is ignored, then $C^g(r_t) = C(r_t)$, and θ^g_d in (19) is reduced to θ^{lat}_d in (10). The estimation of the graph-enhanced background model θ^g_b is in (20), which is exactly parallel to (11) except $C(r_t)$ replaced by $C^g(r_t)$. The language model $P(t|\theta^g_d)$ in (19) for document d is then interpolated with a background model θ^g_b to obtain the smoothed model $\bar{\theta}^g_d$ in (21), which is exactly parallel to (12).

$$P(t|\theta^g_b) = \frac{\sum_{r_t \in \mathcal{C}} C^g(r_t)}{\sum_t \sum_{r_t \in \mathcal{C}} C^g(r_t)}. \quad (20)$$

$$P(t|\bar{\theta}^g_d) = \lambda'_d P(t|\theta^g_d) + (1-\lambda'_d)P(t|\theta^g_b). \quad (21)$$

The language model $\bar{\theta}^g_d$ is in turn used in (1) for ranking.

[3] The initial values do not influence the final results.

III. Spoken Document Expansion with Probabilistic Latent Semantic Analysis

The problem with retrieving documents that are semantically related to the query is that many of these desired documents may not necessarily contain the query term. For example, given the query "airplane", some relevant documents may instead contain the term "aircraft". These relevant documents are given a very small relevance score $S(Q,d)$ in (1) because the unigram language models for the query and the document are very different if they are directly estimated from the term occurrences in the query and the document. This problem can be mitigated by incorporating latent topic analysis approaches. Using such approaches, documents containing the term "aircraft" may be found to belong to the "flying vehicles" latent topic; hence we create a better representation of the document by augmenting the document model with terms related to "flying vehicles" (like "airplane") which are not actually mentioned in the document.

Here we use a very popular approach for latent topic analysis: probabilistic latent semantic analysis (PLSA) [57]. Extension to other latent topic analysis approaches is certainly possible. PLSA uses a set of latent topic variables, $\{T_k, k = 1, 2, ..., K\}$, to characterize the "term-document" co-occurrence relationships in the archive. Given all the spoken documents in the archive, PLSA training yields $P(t|T_k)$, the probability of observing a term t given latent topic T_k, and $P(T_k|d)$, the mixture weight of topic T_k given document d. Based on this latent topic analysis, the probability of observing term t given document d is parameterized as

$$P^T(t|d) = \sum_{k=1}^{K} P(t|T_k)P(T_k|d). \tag{22}$$

The superscript ^{T} indicates it is based on *latent topics*. The parameters $P(t|T_k)$ and $P(T_k|d)$ are learned using the EM algorithm to maximize the following objective function [57]:

$$L = \sum_{d \in \mathcal{C}} \sum_{t} P(t|\theta_d) log P^T(t|d), \tag{23}$$

where θ_d can be θ_d^{1b} in (3) in Section II-A, θ_d^{lat} in (10) in Section II-B or θ_d^{g} in (19) in Section II-C. All the document models in (23) can be unigram language models based on word or subword n-grams (term t can be a word or a subword n-gram), so PLSA models for both words and subword units can all be learned. The maximization of (23) can be understood as the search for a set of parameters $P(t|T_k)$ and $P(T_k|d)$ by minimizing the KL divergence between the document model θ_d and the term distribution $P^T(t|d)$ in (22) obtained from latent topic analysis for all documents d in the collection $\mathcal{C}$.

There are several alternatives to incorporating the above PLSA into the task of information retrieval. One alternative is to project both document and query onto the latent topic space, and rank the documents according to their similarities in terms of latent topic distributions [57]. However, in recent experiments this approach did not always offer satisfactory results [58]. Here, we instead adapt the background model θ_b differently for each document d based on its latent topics, so as to expand the document with semantically related terms via smoothing with the adapted background model [40]. This is realized by interpolating the term distribution for document d based on latent topics, $P^T(t|d)$ in (22), with the general background model θ_b to yield a document-expanded background model $\theta_{b(d)}$ for the document d as in (24), which is document-dependent.

$$P(t|\theta_{b(d)}) = b_d P^T(t|d) + (1 - b_d)P(t|\theta_b). \tag{24}$$

In (24), θ_b can be either θ_b^{1b} in (5) in Section II-A, θ_b^{lat} in (12) in Section II-B, or θ_b^{g} in (21) in Section II-C. Factor b_d can be λ_d in (5) if θ_b^{1b} is used for θ_b, or λ'_d in (12) if θ_b^{lat} or θ_b^{g} are used instead. This document-dependent expanded background model $\theta_{b(d)}$ is then used to smooth the document models θ_d^{1b} in (3) in Section II-A, θ_d^{lat} in (10) in Section II-B, or θ_d^{g} in (19) of Section II-C estimated solely based on the term occurrences in the documents. Therefore, after $\theta_{b(d)}$ smoothing, the probabilities are increased for those words highly related to the topics addressed by the document d in the document model.

IV. Query Expansion with a Query-regularized Mixture Model

Another popular approach for retrieving semantically related documents is query expansion, in which semantically-related terms are automatically added to the query. The expanded query thus enables the retrieval of documents not containing the original query terms but semantically related to the query. Query expansion is often realized using pseudo-relevance feedback (PRF). The top M documents in the first-pass retrieved results with the highest $S(Q,d)$ in (1) are assumed to be relevant (or pseudo-relevant), and the terms that occur frequently in those pseudo-relevant documents are used for query expansion. However, since not all pseudo-relevant documents are truly relevant, and not all words in truly relevant documents are semantically related to the query, it can be difficult to select useful terms for query expansion. Below we borrow the successful query-regularized mixture model [43] from text information retrieval and apply it to spoken content; in addition, we extend it from terms (Section IV-A) to topics (Section IV-B), and from document-level to utterance-level.

A. Term-based Query Expansion

As described and applied on spoken content in Section IV-A1 below, the original query-regularized mixture model is based on terms and is performed at the document level. We further extend this term-based model to the utterance level in Section IV-A2.

1) Document-Level Query Expansion: The query-regularized mixture model assumes that the pseudo-relevant documents are composed of query-related terms and general terms, with a document-dependent ratio between the two. For example, if an irrelevant document is taken as pseudo-relevant, the document's ratio for query-related terms to general ones should be very low. Although these document-dependent ratios and the determination of which terms are query-related are unknown, they can be estimated from the term distributions in the pseudo-relevant documents. Given this estimation, these query-related terms form a new query model θ'_Q, which we use to replace θ_Q in (1).

Suppose the pseudo-relevant spoken document set is $\{d_1, ..., d_m, ..., d_M\}$, where M is the number of documents in the set. With the assumption that the terms in each pseudo-relevant spoken segment are either query-related or are general terms, the document model θ_{d_m} for a pseudo-relevant document d_m should be close to an estimated unigram language model θ'_{d_m} in (25), which is the interpolation of the desired query model θ'_Q to be estimated (for query-related terms) with a background model θ_b (for general terms) with weights α_m and $1 - \alpha_m$.

$$P(t|\theta'_{d_m}) = \alpha_m P(t|\theta'_Q) + (1 - \alpha_m) P(t|\theta_b), \qquad (25)$$

where α_m is the document-dependent interpolation weight for document d_m; this unknown weight must also be estimated. Hence, the goal here is to find the query model θ'_Q and the weights $\alpha_1, \alpha_2, ..., \alpha_M$ for all pseudo-relevant documents $d_1, d_2, ..., d_M$ which minimize $F_1(\theta'_Q, \alpha_1, ..., \alpha_M)$ in (26).

$$F_1(\theta'_Q, \alpha_1, ..., \alpha_M) = \sum_{m=1}^{M} KL(\theta_{d_m} || \theta'_{d_m}), \qquad (26)$$

which is the sum of the KL divergences between θ_{d_m} and θ'_{d_m} for all d_m. The resultant query model θ'_Q then replaces θ_Q in (1). For the spoken documents considered here, the document model θ_{d_m} in (26) can be either $\theta^{1b}_{d_m}$ from the one-best transcriptions in (3) of Section II-A, $\theta^{lat}_{d_m}$ derived from the lattices in (10) of Section II-B, or the graph-enhanced version $\theta^{g}_{d_m}$ in (19) of Section II-C, all of which can be based on either word or subword n-grams. The corresponding background model θ_b is then either θ^{1b}_b in (5) of Section II-A, θ^{lat}_b in (12) of Section II-B, or θ^g_b in (21) of Section II-C. However, the model θ'_Q yielded by minimizing (26) may simply model the common content included in the pseudo-relevant documents; contrary to our desire, it may not necessarily be specifically query-related. In order to handle this problem, θ'_Q is further "regularized" by the original query model θ_Q in (2) based on function $F_2(\theta'_Q)$ as the prior for θ'_Q based on θ_Q:

$$F_2(\theta'_Q) = KL(\theta_Q || \theta'_Q). \qquad (27)$$

Because $F_2(\theta'_Q)$ is the KL divergence between θ_Q and θ'_Q, it is smaller for models θ'_Q that are closer to θ_Q. Therefore, the desired new query model θ'_Q and the weights α_m are estimated by minimizing the objective function

$$F(\theta'_Q, \alpha_1, ..., \alpha_M) = F_1(\theta'_Q, \alpha_1, ..., \alpha_M) + \rho F_2(\theta'_Q), \qquad (28)$$

where parameter ρ controls the influence of the prior function $F_2(\theta'_Q)$. The model θ'_Q estimated via minimizing (28) is not pulled too far away by the pseudo-relevant documents because the function $F_2(\theta'_Q)$ prefers the estimated query model θ'_Q to be similar to the original query model θ_Q. This expanded query model θ'_Q is then used as θ_Q in (1).

The new query model θ'_Q maximizing (28) is obtained using the EM algorithm. Given an initial query model θ^0_Q and a set of initial ratios $\{\alpha^0_1, ..., \alpha^0_M\}$, at the i-th iteration ($i = 1, ..., I$), the probabilities of observing term t in every pseudo-relevant document from the query model θ^{i-1}_Q are computed based on $\{\alpha^{i-1}_1, ..., \alpha^{i-1}_M\}$ in the E step, and then a new query model θ^i_Q and ratios $\{\alpha^i_1, ..., \alpha^i_M\}$ for the i-th iteration are then obtained in the M step. The formulations for the EM algorithm used here are listed as below:

- E step: For each term t in each document in $\{d_1, ..., d_m, ..., d_M\}$, we first compute the posterior probability that the term t is generated from θ^{i-1}_Q:

$$P(\theta^{i-1}_Q | t, d_m) = \frac{\alpha^{i-1}_m P(t|\theta^{i-1}_Q)}{\alpha^{i-1}_m P(t|\theta^{i-1}_Q) + (1 - \alpha^{i-1}_m) P(t|\theta_b)}. \qquad (29)$$

- M step: For each document in $\{d_1, ..., d_m, ..., d_M\}$, we update its ratio of query-related/general terms as

$$\alpha^i_m = \sum_t P(\theta^{i-1}_Q | t, d_m) P(t|\theta_d) \qquad (30)$$

  and the new query model θ^i_Q is updated as

$$P(t|\theta^i_Q) = \frac{\rho P(t|\theta_Q) + \sum_{m=1}^{M} P(t|\theta_d) P(\theta^{i-1}_Q | t, d_m)}{\rho + \sum_t \sum_{m=1}^{M} P(t|\theta_d) P(\theta^{i-1}_Q | t, d_m)}, \qquad (31)$$

  where λ is the parameter in (28).

After I iterations, the model θ^I_Q is finally used as θ'_Q.

2) Utterance-Level Query Expansion: We further extend the above concept from the document level to the utterance level by assuming that *each utterance* in the pseudo-relevant documents has its own ratio of query-related terms to general terms. Let the utterances $\{x_1, ..., x_j, ..., x_J\}$ be the J utterances in the M pseudo-relevant spoken documents $\{d_1, ..., d_m, ..., d_M\}$. The new desired query model θ'_Q with utterance-level query expansion is then obtained by minimizing

$$F'(\theta'_Q, \beta_1, ..., \beta_J) = F'_1(\theta'_Q, \beta_1, ..., \beta_J) + \rho F_2(\theta'_Q), \qquad (32)$$

which is parallel to (28). In (32), $\{\beta_1, ..., \beta_J\}$ is the ratio of query-related terms to general terms for each utterance. $F_2(\theta'_Q)$ is exactly the same as in (27), and

$$F'_1(\theta'_Q, \beta_1, ..., \beta_J) = \sum_{j=1}^{J} KL(\theta_{x_j} || \theta'_{x_j}). \qquad (33)$$

Equation (33) is parallel to (26) except $\{\alpha_1, ..., \alpha_M\}$, θ_{d_m} and θ'_{d_m} in (26) are respectively replaced by $\{\beta_1, ..., \beta_J\}$, θ_{x_j} and θ'_{x_j}. θ'_{x_j} in (33) is also the interpolation of the desired query model θ'_Q (to be estimated) and the background model θ_b parallel to (25) with interpolation weight β_j:

$$P(t|\theta'_{x_j}) = \beta_j P(t|\theta'_Q) + (1 - \beta_j) P(t|\theta_b). \qquad (34)$$

The model θ_{x_j} in (33) is the unigram language model for utterance x_j, which can be based on the one-best transcription $P(t|\theta^{1b}_{x_j}) = \frac{N(t, x_j)}{\sum_t N(t, x_j)}$ parallel to θ^{1b}_d in (3) in Section II-A; or it can be based on lattices $P(t|\theta^{lat}_{x_j}) = \frac{\sum_{r_t \in x_j} C(r_t)}{\sum_t \sum_{r_t \in x_j} C(r_t)}$ parallel to (10) in Section II-B; or it can be based on the graph-enhanced version $P(t|\theta^g_{x_j}) = \frac{\sum_{r_t \in x_j} C^g(r_t)}{\sum_t \sum_{r_t \in x_j} C^g(r_t)}$ parallel to θ^g_d in (19) in Section II-C.

B. Topic-based query expansion

The above query expansion technique is based entirely on terms. Here we further extend this approach to a similar version but base it on latent topics. In topic-based query expansion, everything is in parallel with the term-based query-regularized mixture model as summarized in Subsection IV-A, but here instead of estimating a term-based query model (or query-related term distribution) θ'_Q, or $P(t|\theta'_Q)$, we now seek to estimate a query-related *topic distribution* ϕ_Q over the latent topics, $\{P(T_1|\phi_Q), ..., P(T_k|\phi_Q), ..., P(T_K|\phi_Q)\}$, or $P(T_k|\phi_Q)$, where T_k is a topic, and K is the number of topics. We use ϕ to represent a topic distribution, similar to using θ representing a term distribution. Here we assume the topics T_k are obtained using latent topic analysis such as PLSA, and therefore the probabilities of observing all terms given each latent topic $P(t|T_k)$ are available. Below we use the version of document-level query expansion (over documents d_m) to demonstrate topic-based query expansion. Extension to the utterance level (over utterances x_j) is trivial. For each query Q, the desired topic distribution ϕ_Q is estimated by minimizing the objective function in (35) completely in parallel to (28) or (32).

$$F^T(\phi_Q, \gamma_1, ..., \gamma_M) = F_1^T(\phi_Q, \gamma_1, ..., \gamma_M) + \rho F_2(\phi_Q), \quad (35)$$

where

$$F_1^T(\phi_Q, \gamma_1, ..., \gamma_M) = \sum_{m=1}^{M} KL(\theta_{d_m} || \theta^T_{d_m}), \quad (36)$$

$F_1^T(\phi_Q, \gamma_1, ..., \gamma_M)$ in (36) are exactly in parallel to (26), except that α_m is replaced by γ_m, and the term distribution θ'_{d_m} in (25) is replaced by $\theta^T_{d_m}$.

$$P(t|\theta^T_{d_m}) = \gamma_m P^T(t|\phi_Q) + (1 - \gamma_m) P(t|\theta_b), \quad (37)$$

where

$$P^T(t|\phi_Q) = \sum_{k=1}^{K} P(t|T_k) P(T_k|\phi_Q). \quad (38)$$

Equation (38) is in parallel with (22): while (22) is for document d, (38) is for topic distribution ϕ_Q. $F_2(\phi_Q)$ in (35) is exactly (27).

After obtaining the topic distribution ϕ_Q by minimizing the objective function in (35), the term distribution given the desired expanded query model θ^T_Q estimated based on the latent topics is

$$P^T(t|\theta^T_Q) = \sum_{k=1}^{K} P(t|T_k) P(T_k|\phi_Q). \quad (39)$$

Equation (39) is exactly the same as (38), except we replace ϕ_Q in (38) by θ^T_Q in (39) because this is a term distribution rather than a topic distribution, and is thereby better expressed as θ^T_Q as in (39) rather than ϕ_Q in (38). The superscript ^{T} in θ^T_Q indicates this query model (or term distribution) is based on latent topics. This term distribution or query model θ^T_Q can be further interpolated with the word-based expanded query model θ'_Q obtained with (28) or (32) to yield a query model θ''_Q considering both words and topics:

$$P(t|\theta''_Q) = \delta P(t|\theta'_Q) + (1 - \delta) P^T(t|\theta^T_Q), \quad (40)$$

where δ is an interpolation weight. The expanded query model ϕ_Q obtained in (35) but expressed as θ^T_Q in (39) based on latent topics, or the interpolated version θ''_Q in (40) can then be used as θ_Q in (1).

V. Experimental Setup

In the experiments, we used a Mandarin Chinese broadcast news corpus as the spoken document collection $\mathcal{C}$ to be retrieved through [59][4]. The news stories were recorded from radio or TV stations in Taipei from 2001 to 2003. There were a total of 5047 news stories, with a total length of 198 hours. The story lengths ranged from 68 to 2934 characters, with an average of 411 characters per story. The average time duration per news story was 2.35 minutes. One hundred sixty-three text queries and their relevant spoken documents (not necessarily including the queries) were provided by 22 graduate students. The query lengths ranged from 1 to 4 Chinese words with an average of 1.6 words, or 1 to 8 Chinese characters with an average of 2.7 characters. Some examples for the queries are "typhoon disaster (颱風災情)", "election (選舉)" and "important stock market information (股市重大訊息)"[5]. The number of relevant documents for each query ranged from 1 to 50 with an average of 19.5. Forty-one out of 163 queries were used in the development set for parameter tuning, while the remaining 122 queries were testing queries.

In order to evaluate the retrieval performance of the proposed approaches with respect to different recognition conditions, we used different acoustic and language models to transcribe the spoken documents. As listed below, we used two different recognition conditions to generate the lattices for the spoken document collection $\mathcal{C}$:

- Condition (I): We used a tri-gram language model trained on 39M words of Yahoo news, and a set of acoustic models with 48 Gaussian mixtures per state and 3 states per model trained on a training corpus of 24.5 hours of Mandarin broadcast news different from the above mentioned collection tested here. One hundred forty-seven right context-dependent Initial models plus context-independent Final models were used as the acoustic models. Here Initial is the initial consonant of a Mandarin syllable, and Final is the vowel/diphthong part with an optional medial or nasal ending. This kind of acoustic model has been heavily used to recognize Mandarin speech. The acoustic features used were MFCCs with cepstral mean and variance normalization (CMVN). The character accuracy for the whole collection was 54.43%.
- Condition (II): We cascaded perceptual linear predictive (PLP) features with Mandarin phoneme posterior probabilities estimated by a multilayer perceptron (MLP) trained on 10 hours of Mandarin broadcast news different from those tested here as the input features for a Tandem system. A tri-gram language model trained on 98.5M words of news from several different sources, and a set of acoustic models with 48 Gaussian mixtures per state

[4] Publicly available via the Association for Computational Linguistics and Chinese Language Processing (http://www.aclclp.org.tw).

[5] All queries were in Mandarin Chinese, but translated into English.

and 3 states per model trained on the above training set of 24.5 hours of broadcast news were used. The same configuration of 147 right context-dependent Initial models plus context-independent Final models used for Condition (I) was used as well. The character accuracy was 62.13%.

Both conditions (I) and (II) used a 60K-word lexicon, and the beam width for decoding was 100. Since 48% and 31% of the speech in the corpus were produced by the reporters and interviewees respectively which were highly spontaneous including relatively high background noise, the character accuracy for them was relatively low (excluding the speech produced by the reporters and interviewees, the character accuracies for conditions (I) and (II) were respectively 65.00% and 74.96%). Such relatively low recognition accuracies are realistic in the real world and mirror situations where the graph-based approaches proposed here may be helpful, since retrieval performance inevitably depends on recognition accuracy. After the recognition systems transcribed each utterance into a word lattice, we further transformed each Chinese word arc in the lattice into a sequence of concatenated corresponding Chinese character and Mandarin syllable arcs to respectively form character and syllable lattices, that is, characters and syllables were taken as the subword units here in the experiments. Therefore, for each utterance, there were three lattices: word-, character-, and syllable-based, for the above two recognition Conditions (I) and (II).

We used mean average precision (MAP) as the evaluation measure for the following experiments [60]. The pair-wise t-test with a significance level of 0.05 was used to gauge the significance of performance improvements. For the language modeling retrieval approach, the parameter κ below (5) and (12) was 1000 in all the experiments[6]. Frame-based DTW was used to compute distance $d(r_i, r_j)$ in (15). The acoustic features used in speech recognition for the spoken documents were also used in frame-based DTW (that is, MFCC with CMVN for Condition (I), and PLP plus phoneme posterior probabilities for Condition (II)), and Euclidean distance was taken as the distance measure between two acoustic feature vectors. A slope constraint was used in the frame-based DTW to handle speaking rate distortion. With the slope constraint, a feature sequence can match another sequence whose length is in a range of $1/\mu$ to μ times the length of the former; μ was set to 3 in the following experiments [62], [63]. The distance obtained by the above frame-based DTW in the reference, $d'(r_i, r_j)$, is asymmetric; $d'(r_i, r_j) \neq d'(r_j, r_i)$. Here $d(r_i, r_j)$ in (15) is the average of $d'(r_i, r_j)$ and $d'(r_j, r_i)$. For the graph construction in Section II-C, nodes r_i and r_j were connected if r_i was among the k-nearest neighbors of r_j based on $S(r_i, r_j)$ in (15), and if r_j was among the k-nearest neighbors of r_i ($k = 10$ in the experiments). α in (16) was set to 0.5. The number of iterations I for the EM algorithm in Section IV-A1 was 10 in all the experiments. Unless otherwise specified, all the other parameters were determined using the development set. That is, the optimal values for the parameters were based on the development set, and the same values were applied on the testing set.

VI. Experimental Results

A. Initial Baselines for Language Modeling Retrieval Approach

TABLE I: Comparison of MAP performance for the testing queries yielded by one-best transcriptions (row (1)) and lattices (row (2)) on Conditions (I) and (II) (Parts (a) and (b)) with Okapi BM25 (columns labeled "BM25") or basic language modeling retrieval approach (columns labeled "LM"). Both retrieval approaches were based on word unigrams without document/query expansion. The superscripts * in row (2) indicate performance significantly better than the corresponding values in row (1), and the superscripts † in the columns labeled "LM" indicate performance significantly better than those labeled "BM25" in the same parts.

	(a) **Condition (I)**		(b) **Condition (II)**	
	BM25	LM	BM25	LM
(1) *One-best*	42.46%	44.91%†	46.42%	49.59%†
(2) *Lattice*	44.79%*	45.68%*†	48.26%*	50.70%*†

We should first justify the use of the language modeling retrieval approach as the baseline in this study. Table I compares the results of the testing queries with Okapi BM25 [64] (columns labeled "BM25"), and language modeling retrieval approach (columns labeled "LM") without document/query expansion. Okapi BM25 is a standard retrieval approach based on term frequencies, inverse document frequencies, and document lengths equally useful for text or spoken content retrieval. All parameters in Okapi BM25 were tuned on the development set. All results in Table I were based on word unigrams only; words were used as terms here. Parts (a) and (b) are respectively the results for Conditions (I) and (II). Row (1) is the results based on one-best transcriptions. That is, we computed the term frequencies, inverse document frequencies, and document lengths based on the one-best transcriptions as in the text form for Okapi BM25. For the language modeling retrieval approach, $\bar{\theta}_d^{1b}$ in (5) of Section II-A was taken as θ_d in (1). Row (2) is for lattices. For BM25 in row (2), the expected term frequencies and expected document lengths computed based on the lattices were used, but the inverse document frequencies were computed based on the one-best transcriptions[7]. For the language modeling retrieval approach in row (2), $\bar{\theta}_d^{lat}$ in (12) in Section II-B was used in (1). The superscripts * in row (2) indicate performance significantly better than the corresponding values in row (1), while the superscripts † in the columns labeled "LM" indicate

[6] This is also the default value of the Lemur Toolkit [61] for the same smoothing approach.

[7] To compute the inverse document frequency of a term, we need the number of documents containing the term. In text retrieval, this number is obtained in a straightforward way; the same approach is used for the one-best transcriptions. However, to compute this number on the lattices, we take as containing the term only those documents with expected frequencies of the term that exceed a threshold. Intuitively, there seems to be no direct way to decide this threshold. Therefore, for simplicity, we here only compute inverse document frequencies on the one-best transcriptions. There are more sophisticated approaches for estimating the inverse document frequencies based on lattices with training data [52], but out of the scope of this paper.

performance significantly better than those labeled “BM25” in the same parts.

Clearly, the lattice-based results significantly outperformed those based on one-best transcriptions under both Conditions (I) and (II), regardless of the retrieval approach (rows (2) vs (1)). Table I also shows that the language modeling retrieval approach is significantly better than Okapi BM25 for the experimental setup here. Hence, we chose the language modeling retrieval approach as our basic retrieval approach here[8]. With the same retrieval approach, Condition (II) always yielded much better results than Condition (I) whether one-best transcriptions or lattices were considered; retrieval performance depends on recognition accuracy. The MAP yielded by manual transcriptions for the language modeling retrieval approach here was 62.16%, which was much better than the results for recognition condition (II) with lattices and language modeling retrieval approach (62.16% vs 50.70% in Table I). This observation shows that even given the use of lattices, recognition errors still led to degraded retrieval performance. Hence, even with lattices, improved techniques are called for.

B. Presentation of Overall Experimental Results

Tables II and III respectively show the overall experimental results for the testing queries based on Conditions (I) and (II), each with the upper half (Part (A)) for results directly from lattices ($\bar{\theta}_d^{lat}$ in (12) was used) and the lower half (Part (B)) for the proposed graph-enhanced version ($\bar{\theta}_d^{g}$ in (21) was used). For each half, rows (1), (2), and (3) are respectively for word unigram, character bi-gram, and syllable bi-gram language models, that is, the terms t used were words, character bi-grams, and syllable bi-grams[9]. We further weighted the relevance scores $S(Q, d)$ in (1) obtained in rows (1) to (3) and then summed them up to obtain the results in row (4). The weights for each unit, or each row, were determined using the development set.

Columns (a), (b), and (c) in Tables II and III respectively list results for the basic language modeling retrieval approach without any expansion, with document expansion, and with utterance-level query expansion. The columns (c-1) and (c-2) are respectively for term only and for term plus topic. Column (d) then integrated document expansion with query expansion (term plus topic). Superscripts $^{\alpha}$, $^{\beta}$, $^{\gamma}$, and $^{\delta}$ respectively indicate performance significantly better than the corresponding results in columns (a), (b), (c-1), and (c-2) in the same row. Due to limited computational resources, we only constructed the graphs respectively for the 30,000 words, character bi-grams, and syllable bi-grams with the highest tf-idf scores computed on the 1-best transcriptions. For those terms without graphs, we simply set $C^g(r_t)$ equal to $C(r_t)$ in (19) when computing θ_d^g. In both Tables II and III, the superscripts * in the lower halves (Parts (B)) indicate performance significantly better than the corresponding results in the upper halves (Parts (A)). Although we only enhanced the top 30,000 terms with graphs, we still observe encouraging results in the experiments.

C. Discussion on the Basic Language Modeling Approach without Document/Query Expansion

We now focus on columns (a) of Tables II and III for the basic language modeling retrieval approach without any document or query expansion. We found that the proposed graph-based enhancement approach always improved the retrieval performance significantly (parts (B) vs (A) in columns (a) of Tables II and III). Because the score propagation over the graphs brought the document language models closer to the true term distributions in speech, the graph-enhanced language models led to better results. We also observed that in columns (a) the results for language models based on character bi-grams were better than those based on words and syllable bi-grams (rows (2) vs (1), (3) in columns (a) of Tables II and III). Language models based on character bi-grams outperformed words because they handled OOV queries better. For example, some longer OOV words cannot be correctly recognized, but part may be correctly transcribed into correct character bi-grams. Although syllables are also very helpful subword units (each character is produced as a monosyllable in Mandarin Chinese), in Mandarin Chinese different characters with different meanings often correspond to the same syllable (there are far fewer syllables than there are characters). Hence, syllable bi-grams were not as discriminative as character bi-grams in representing the semantics in the documents, even though syllable bi-grams also mitigated the OOV problem somewhat. Moreover, since syllables were not as discriminative as words and characters, they were more susceptible to recognition errors (an erroneous syllable may lead to a whole group of erroneous characters and more than one erroneous word). Therefore, with higher recognition accuracy (Condition (II) of Table III), syllable bi-grams and words were comparable (rows (3) vs (1) in column (a) of Table III), whereas given poor recognition accuracy (Condition (I) of Table II), syllable bi-grams were the worst among the three types of terms (rows (3) vs (1), (2) in column (a) of Table II). Furthermore, although characters are more discriminative than syllables, syllables usually have higher accuracies because incorrectly recognized characters often correspond to correct syllables. This is why words, character bi-grams, and syllable bi-grams carry complementary information. Hence, their integration outperformed them all individually (rows (4) vs (1), (2), (3) in columns (a) of Tables II and III).

We then analyse the estimation accuracy for θ_d^{lat} and θ_d^{g} with respect to the reference model θ_d^{ref}. The reference model θ_d^{ref} is from the manual transcriptions, where

$$P(t|\theta_d^{ref}) = \frac{N_c(t,d)}{\sum_t N_c(t,d)}, \tag{41}$$

where $N_c(t,d)$ is the occurrence counts of the term t in the manual transcriptions of spoken document d. The results of the first one hundred spoken documents d in the spoken archive

[8] Our intent here is not to claim that the language modeling retrieval approach is better than Okapi BM25. Although recent work shows that the language modeling retrieval approach outperforms Okapi BM25 in both text and speech retrieval [46], [65], evidence that the language modeling retrieval approach is superior to BM25 is as yet insufficient.

[9] Word bi-gram was not helpful (even integrated with other units), so we do not report the results.

TABLE II: Condition (I) MAP performance for the testing queries, with upper (Part (A)) and lower (Part (B)) halves respectively for results directly from the lattices without and with graph-enhancement. Columns (a), (b), and (c) are respectively for the basic language modeling retrieval approach without any document or query expansion, with document expansion, and with utterance-level query expansion. Columns (c-1) and (c-2) are respectively for terms only and for terms plus topics. Document expansion (column (b)) and query expansion including topic information (column (c-2)) was then integrated in column (d). Superscripts $^{\alpha}$, $^{\beta}$, $^{\gamma}$, and $^{\delta}$ respectively indicate performance significantly better than the corresponding results in columns (a), (b), (c-1), and (c-2) in the same row. Superscript * indicates that the results in Part (B) are significantly better than the corresponding results in part (A). Rows (1), (2), and (3) are respectively for word unigram, character bi-gram, and syllable bi-gram language models, that is, the terms t used were words, character bi-grams, and syllable bi-grams respectively in rows (1), (2), and (3). A weighted sum of the relevance scores obtained in rows (1) to (3) yielded the results in row (4), with the weights determined on the development set.

Condition (I)		(a) Basic LM	(b) Document expansion	(c) Query expansion (utterance-level)		(d) Document plus query expansion (b) + (c-2)
				(c-1) *Term-based*	(c-2) *Term+topic-based*	
(A) Lattice	(1): Word 1-gram	45.68%	48.26%$^{\alpha}$	48.58%$^{\alpha}$	48.91%$^{\alpha}$	49.98%$^{\alpha\beta\gamma\delta}$
	(2): Char 2-gram	45.99%	47.02%$^{\alpha}$	46.94%	46.96%	48.11%$^{\alpha\beta}$
	(3): Syl 2-gram	43.32%	45.52%$^{\alpha}$	44.94%$^{\alpha}$	45.23%$^{\alpha\gamma}$	46.89%$^{\alpha\beta\gamma\delta}$
	(4): (1)+(2)+(3)	48.46%	51.29%$^{\alpha}$	50.12%	49.63%	52.54%$^{\alpha\beta\gamma\delta}$
(B) Graph-enhanced	(1): Word 1-gram	47.42%*	50.01%$^{\alpha}$	49.10%$^{*\alpha}$	50.48%$^{*\alpha\gamma}$	50.51%$^{\alpha\beta\gamma}$
	(2): Char 2-gram	48.18%*	48.88%$^{*\alpha}$	48.19%*	48.21%$^{*\gamma}$	49.16%$^{*\alpha}$
	(3): Syl 2-gram	44.99%*	46.79%$^{*\alpha}$	46.11%$^{*\alpha}$	46.55%$^{*\alpha\gamma}$	47.41%$^{\alpha\beta\gamma\delta}$
	(4): (1)+(2)+(3)	50.38%*	52.70%$^{*\alpha}$	51.10%*	52.71%$^{*\alpha\gamma}$	54.00%$^{*\alpha\beta\gamma\delta}$

TABLE III: Condition (II) MAP performance for the testing queries in Table II.

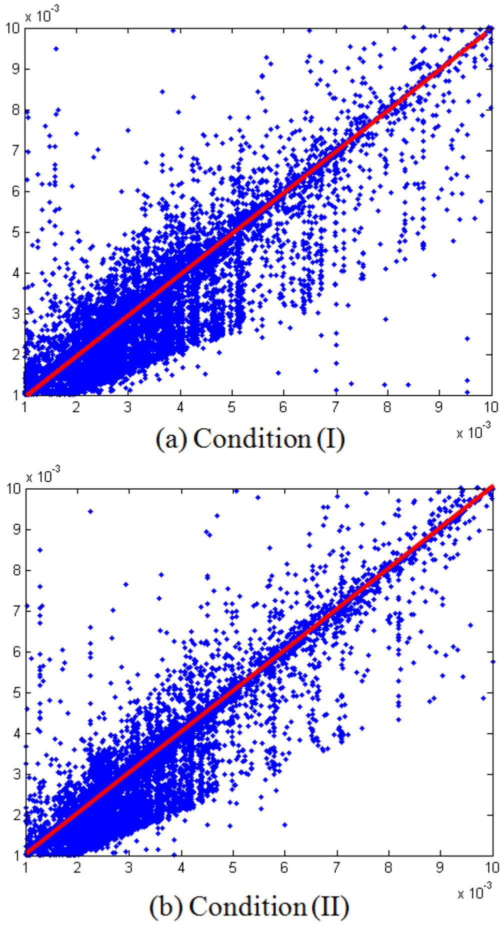
Condition (II)		(a) Basic LM	(b) Document expansion	(c) Query expansion (utterance-level)		(d) Document plus query expansion (b) + (c-2)
				(c-1) *Term-based*	(c-2) *Term+topic-based*	
(A) Lattice	(1): Word 1-gram	50.70%	53.29%$^{\alpha}$	52.76%$^{\alpha}$	53.09%$^{\alpha\gamma}$	54.41%$^{\alpha\beta\gamma\delta}$
	(2): Char 2-gram	52.55%	53.58%$^{\alpha}$	53.10%	53.20%	54.83%$^{\alpha\beta\gamma\delta}$
	(3): Syl 2-gram	50.58%	52.05%$^{\alpha}$	50.98%	50.96%	52.62%$^{\alpha\beta\gamma\delta}$
	(4): (1)+(2)+(3)	55.11%	56.66%$^{\alpha\gamma}$	55.22%	55.65%$^{\gamma}$	56.94%$^{\alpha\gamma\delta}$
(B) Graph-enhanced	(1): Word 1-gram	52.05%*	53.92%$^{*\alpha}$	53.91%$^{*\alpha}$	54.08%$^{*\alpha}$	54.98%$^{\alpha\beta\gamma\delta}$
	(2): Char 2-gram	53.55%*	54.50%$^{*\alpha}$	54.31%$^{*\alpha}$	54.37%$^{*\alpha}$	55.15%$^{\alpha\beta}$
	(3): Syl 2-gram	51.45%*	52.88%$^{\alpha}$	52.10%	52.62%$^{*\alpha\gamma}$	53.24%$^{\alpha\gamma\delta}$
	(4): (1)+(2)+(3)	55.32%	57.23%$^{\alpha}$	56.04%	56.60%$^{*\gamma}$	57.97%$^{*\alpha\beta\gamma\delta}$

on Conditions (I) and (II) are respectively plotted in Figs. 3 (a) and (b). In Fig. 3, each point represents a word t in a document d (so we have a total of V times one hundred points in each figure, where V is the vocabulary size). The x scale of Fig. 3 is the absolute value of the difference between $P(t|\theta_d^{lat})$ and $P(t|\theta_d^{ref})$, or $|P(t|\theta_d^{lat}) - P(t|\theta_d^{ref})|$ ($|x|$ means the absolute value of x), while the y scale is the absolute value of difference between $P(t|\theta_d^{g})$ and $P(t|\theta_d^{ref})$, or $|P(t|\theta_d^{g}) - P(t|\theta_d^{ref})|$. The red line is the line of $x = y$, that is, points for which the estimation errors of $P(t|\theta_d^{lat})$ and $P(t|\theta_d^{g})$ with respect to $P(t|\theta_d^{ref})$ are identical. The points below the red line are those having $|P(t|\theta_d^{lat}) - P(t|\theta_d^{ref})| > |P(t|\theta_d^{g}) - P(t|\theta_d^{ref})|$, or the graph-based enhancement approach yielded better estimation than lattices; while the points above the red line indicate on other way. From Figs. 3 (a) and (b), it is clear that there are much more points below the red lines in both figures. This shows that the graph-enhancement approach usually provided more accurate estimation under both conditions.

TABLE IV: KL divergence values for the estimated document language models, either directly from lattices (Part (A)) or with graph-enhancement (Part (B)), evaluated against the correct document models, with rows (1), (2), and (3) respectively for word unigrams, character bi-grams, and syllable bi-grams.

KL divergence	(A) **Lattice**		(B) **Graph-enhanced**	
	Condition (I)	Condition (II)	Condition (I)	Condition (II)
(1) Word 1-gram	4.11	3.19	4.06	3.15
(2) Char 2-gram	6.27	4.57	6.20	4.52
(3) Syl 2-gram	3.47	2.91	3.41	2.80

Table IV shows the KL divergence values for the document language models obtained directly from lattices (with $\bar{\theta}_d^{lat}$ in (12) in Part (A)) and those enhanced with graphs (with $\bar{\theta}_d^{g}$ in (21) in Part (B)), evaluated against the reference document models θ_d^{ref} based on manual transcriptions. The values in Parts (A) and (B) in Table IV are respectively the average values of $KL(\theta_d^{ref}||\bar{\theta}_d^{lat})$ and $KL(\theta_d^{ref}||\bar{\theta}_d^{g})$ for all documents. Smaller KL divergence values imply the document models



(a) Condition (I)

(b) Condition (II)

Fig. 3: *The estimation accuracy of $P(t|\theta_d^{lat})$ and $P(t|\theta_d^{g})$ with respect to $P(t|\theta_d^{ref})$, where θ_d^{ref} is the reference model from the manual transcriptions. (a) and (b) are respectively for Conditions (I) and (II). Each point in the figures is for a word t in the first one hundred spoken documents d in the spoken archive.* ***The estimation error of lattice-based language models $|P(t|\theta_d^{lat}) - P(t|\theta_d^{ref})|$ is the x scale, while that with graph-based enhancement approach $|P(t|\theta_d^{g}) - P(t|\theta_d^{ref})|$ is the y scale.*** *The red line is the line of $x = y$.*

estimated from spoken documents were closer to those based on manual transcriptions. We found that the graph-enhanced language models $\bar{\theta}_d^{g}$ always yielded smaller KL divergence values than the original lattice-based models $\bar{\theta}_d^{lat}$ (Parts (B) vs (A)). This verifies that the proposed approach indeed helped to mitigate the problem of recognition errors and brought the estimated document language models closer to the correct term distributions, and explains why the proposed graph-based enhancement approach yielded improvements.

D. Document Expansion

Columns (b) in Tables II and III list the results for document expansion on the testing set, in which the document models were smoothed using the document-expanded background model $\theta_{b(d)}$ in (24) of Section III, but with θ_d^{lat} in (10) of Section II-B and θ_d^{g} in (19) of Section II-C respectively interpolated with $\theta_{b(d)}$ in Parts (A) and (B). Here the PLSA models were learned from the one-best transcriptions, that is, θ_d^{1b} in (3) was taken as θ_d in (23) of Section III. The number of PLSA latent topics (K in (22) in Section III) was determined using the development set. The tables clearly show that PLSA-based document expansion significantly improved the retrieval performance for both word and subword-based language models and under different recognition conditions (columns (b) vs (a) in Tables II and III).

When comparing Parts (A) and (B) of columns (b) in Tables II and III, we note that the proposed graph-based enhancement approach yielded additional improvements for document expansion in all cases (Parts (B) vs (A) in columns (b) in Tables II and III). Graph score propagation improved the scores of the hypothesized regions, yielding better term distributions for θ_b^{g} than θ_b^{lat} to be used as θ_b when estimating $\theta_{b(d)}$ in (24) of Section III. This better background model $\theta_{b(d)}$ is then further used to smooth the better document model θ_d^{g} than the original model θ_d^{lat}, which led to the improvements observed here. Moreover, we found that integrating the results of word-, character-, and syllable-based language models always improved the results for document expansion whether or not the graph-based approach was applied (rows (4) vs (1), (2), (3) in Parts (A) and (B) of columns (b) in Tables II and III).

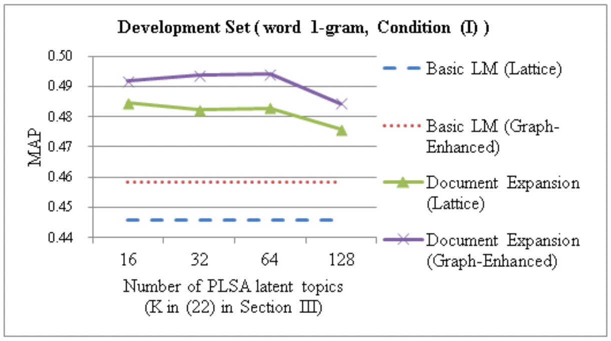


Fig. 4: An example illustrating the influence of different numbers of PLSA latent topics ($K = 16, 32, 64, 128$) for document expansion on the development set. The results in this figure were based on word unigrams; the lattices were generated under Condition (I).

Fig. 4 is an example illustrating the influence of the numbers of PLSA latent topics ($K = 16, 32, 64, 128$) for document expansion on the development set. Only the results based on word unigrams with lattices generated in Condition (I) were reported in Fig. 4. Similar phenomena were observed in all other cases. The curves labeled “Basic LM” and “Document Expansion” are respectively the results without and with document expansion. Since “Basic LM” did not utilize latent topics, its results did not depend on the number of PLSA latent topics. The curves labeled “Lattice” and “Graph-Enhanced” were respectively the results based on lattices and further enhanced by the graph-based approach. We can see from this figure that document expansion improved retrieval performance regardless of the number of latent topics (“Document Expansion” vs “Basic LM” in Fig. 4). Taking the results without and with the graph-based enhancement approach, we also see the proposed approach improved the performance regardless of the number of latent topics (“Lattice” vs “Graph-Enhanced” in Fig. 4). We also find that a too large number of latent topics ($K = 128$) yielded lower improvements from document expansion. Obviously because when K is too large, some semantically related terms may be belonged to different latent topics.

E. Query Expansion

TABLE V: MAP performance yielded by document-level (column (b-1)) and utterance-level (column (b-2)) term-based query expansion on the testing set compared with the basic language modeling approach (column (a)), all with word unigram only for Conditions (I) and (II). The superscripts $^{\alpha}$ and $^{\beta}$ respectively indicate performance significantly better than the corresponding results in columns (a) and (b-1).

	(a) Basic LM	(b) Term-based query expansion	
		(b-1) Document-level	(b-2) Utterance-level
Condition (I)	45.68%	46.46%$^{\alpha}$	48.58%$^{\alpha\beta}$
Condition (II)	50.70%	51.19%$^{\alpha}$	52.76%$^{\alpha\beta}$

For query expansion, we first compare the document-level and utterance-level term-based query expansion on the testing set using the language models obtained from the original lattices for word unigram only. The number of pseudo-relevant documents M in Section IV and the parameter ρ in (28) and (32) were all determined using the development set. Table V reports the MAP results. Column (a) is the results of the basic language modeling approach (same as rows (1) of Parts (A) and columns (a) in Tables II and III). Columns (b-1) and (b-2) are for term-based query expansion in Section IV-A. The pseudo-relevant documents used for query expansion were taken from the first-pass results based on $\bar{\theta}_d^{lat}$ in (12) of Section II-B (those for column (a) of this table). Columns (b-1) and (b-2) are respectively for term-based query expansion on document and utterance levels as in Section IV-A1 and IV-A2. In column (b-1), the expanded new query model θ_Q' was obtained in (28), and θ_d^{lat} in (10) and θ_b^{lat} in (11) were taken as θ_{d_m} in (26) and θ_b in (25); in column (b-2), θ_Q' was obtained by (32), and θ_b in (34) was still θ_b^{lat}, but θ_{x_j} in (33) was the language model for utterance x_j. The superscripts $^{\alpha}$ and $^{\beta}$ in columns (b-1) and (b-2) of Table V respectively indicate performance significantly better than columns (a) (basic language modeling approach) and (b-1) (document-level query expansion). We found both document-level and utterance-level term-based query expansion significantly outperformed the baselines (columns (b-1), (b-2) vs (a)), and the utterance-level approach extended in this paper was significantly better than document-level in all cases (columns (b-2) vs (b-1)). Because it is natural that different utterances in the same document have different degrees of relevance with respect to the queries, the fact that utterance-level query expansion offers better results is intuitive. Since utterance-level query expansion is better, in the following discussions, only utterance-level query expression is considered.

Now we return to Tables II and III but concentrate on columns (c-1) for utterance-level term-based query expansion. The results using the basic language modeling retrieval approach in columns (a) in the same row of the same table were taken as the first-pass results for defining the pseudo-relevant documents for query expansion. Columns (c-1) are for term-based θ_Q' estimated by (32). We found that utterance-level term-based query expansion outperformed the baseline language modeling retrieval approach without expansion in all cases (columns (c-1) vs (a) in Tables II and III). We also observe that query expansion was not very effective in some cases (for example, column (c-1) in row (2) in part (B) of Table II). Because the query models were estimated from pseudo-relevant documents corrupted with recognition errors, this estimation did not necessarily provide high quality query models.

Now we compare Parts (A) and (B) for columns (c-1) in Tables II and III for query expansion without and with graph-based enhancement. Here the expanded query model θ_Q' and θ_Q'' for columns (c-1) were matched against $\bar{\theta}_d^{lat}$ in (12) of Section II-B and $\bar{\theta}_d^{g}$ in (21) of Section II-C respectively in Parts (A) and (B). The graph-based enhancement improved query expansion in all cases (parts (B) vs (A) in columns (c-1) in Tables II and III). The improvements were achieved by two factors. First, here the results of columns (a) in the same rows were taken as the first-pass retrieved results, which improved with graph-based enhancement and thereby included more relevant documents in the pseudo-relevant document set. Second, when matching the expanded query model θ_Q' against the document models, the document model $\bar{\theta}_d^{g}$ obtained with graph-based enhancement was better than $\bar{\theta}_d^{lat}$.

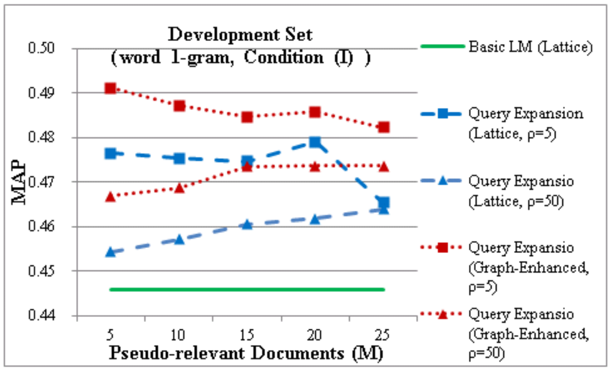


Fig. 5: An example (word unigram under condition (I)) utterance-level term-based query expansion for different numbers of pseudo-relevant documents M ($M = 5, 10, 15, 20, 25$) and different values of ρ in (32) ($\rho = 5, 50$) on the development set. The dashed and dotted curves are respectively the results without and with graph-based enhancement.

Fig. 5 is an example illustrating the influence of the number of pseudo-relevant documents M and the value of parameter ρ in (32) for utterance-level term-based query expansion on the development set with word unigrams in Condition (I). Similar phenomena were also observed in all other cases. The solid line (labeled “Basic LM”) is the results without query expansion. The dashed curves (labeled “Lattice”) are the results of utterance-level term-based query expansion based on lattices, while the dotted curves (labeled “Graph-Enhanced”) are those with graph-based enhancement, all as functions of the parameter M. Parameter ρ controlled the influence of the prior function; a smaller ρ means the expanded queries were less similar to the original queries, or the M pseudo-relevant documents had more influence on the expanded queries. In Fig. 5, a smaller ρ ($\rho = 5$) yielded better results, but performance decreased seriously for large values of M ($M = 25$). Because larger values of M led to more documents in the first-pass retrieval results (some of which may not be relevant) were considered as pseudo-relevant in the query expansion process, for smaller values of ρ, these noisy pseudo-relevant documents had more influence, thus degrading the performance of query expansion. When ρ was large ($\rho = 50$), however, the results were insensitive to M, but since the expanded queries were more similar to the original ones, query expansion yielded only limited improvements. On the other hand, it is clear that given fixed values for M and ρ, the proposed graph-based enhancement approach improved the performance of query expansion in all cases.

F. Topic-based Query Expansion

Columns (c-2) in Tables II and III are for integrated term-based and topic-based query expansion in Section IV-B, or with θ''_Q in (40). Parameter δ in (40) was decided using the development set. We did not tune the number of PLSA latent topics K, the number of pseudo-relevant documents M, or parameter ρ in (35) for topic-based query expansion. Instead, we set the values of K to be the same values used in the document expansion (column (b)) of the same row in Tables II and III, and the values of M and ρ to be the same values used in the term-based query expansion (column (c-1)). Comparing columns (c-2) and (c-1), we find that using topic-based query expansion in addition (θ''_Q in (40) in Section IV-B) offered extra improvements as compared to the original term-based version (θ'_Q in Section IV-A), with two exceptions: columns (c-2) vs (c-1) in row (4) in part (A) of Table II and in row (3) in part (A) of Table III. Although the improvements obtained were not large, most were significant (those with superscript $^{\gamma}$). This implies that the latent topic information was helpful for query expansion in most cases, although it can induce undesired noisy terms.

G. Query Expansion plus Document Expansion

Since document and query expansion use different mechanisms to take into account semantics, we were interested to see if their improvements were additive. The results are in columns (d) of Tables II and III, for which both document and query models in (1) were expanded. In the experiments, the results in columns (b) with document expansion were taken as the first pass to obtain the pseudo-relevant documents, which included more relevant documents than the results in columns (a) because of the help of document expansion, and the estimated new query model θ''_Q (in (40) in Section IV-B) was matched against the document model smoothed by the document-expanded background model $\theta_{b(d)}$. The values of all parameters were set to be the same values used in the document expansion (column (b)), term-based query expansion (column (c-1)), and topic-based query expansion (column (c-2)) of the same row in Tables II and III. From columns (d), we find that applying document and query expansion jointly outperformed the individual approaches (columns (d) vs (b), (c-1), (c-2)). In addition, the proposed graph-based enhancement also yielded extra improvements even with the joint application of query and document expansion (Parts (B) vs (A) in columns (d) in Tables II and III).

VII. Conclusion

To improve the semantic retrieval of spoken content, we here proposed using acoustic similarity graphs to estimate more accurate term distributions for language modeling for the spoken documents. The spoken document language models thus enhanced were then applied on the language modeling retrieval approach, document expansion, and various versions of query expansion. Improved performance for the proposed approach was observed in two different recognition conditions on a corpus of broadcast news in Mandarin Chinese. The proposed approaches were also shown to be equally applied to document models based on different term granularities including words, character bi-grams, and syllable bi-grams for Mandarin Chinese. Moreover, both document expansion based on topic analysis and query expansion based on the query-regularized mixture model were shown helpful, and information from different term granularities were fused to offer better performance. Finally, by integrating the proposed approach with all the techniques, including document expansion, query expansion, and the fusion of the information from different term granularities, we achieved an improvement of 20.2% relative over the baseline using one-best word sequences (from 44.91% to 54.00% in terms of MAP) for the lower character accuracy condition (Condition (I) in Section V), and 16.9% relative over the baseline (from 49.59% to 57.97%) for the higher accuracy condition (Condition (II) in Section V).

References

[1] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, pp. 1–19, 2006.

[2] C. Chelba, T. J. Hazen, and M. Saraçlar, "Retrieval and browsing of spoken content," in *IEEE Signal Processing Magazine 25(3), pp. 39-49*, 2008.

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

[4] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," in *HLT-NAACL*, 2004.

[5] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Comput. Speech Lang.*, vol. 21, pp. 458–478, 2007.

[6] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *ACL*, 2005.

[7] Y.-C. Pan, H.-L. Chang, and L.-S. Lee, "Analytical comparison between position specific posterior lattices and confusion networks based on words and subword units for spoken document indexing," in *ASRU*, 2007.

[8] T. Hori, I. L. Hetherington, T. J. Hazen, and J. Glass, "Open vocabulary spoken utterance retrieval using confusion networks," in *ICASSP*, 2007.

[9] C. Allauzen, M. Mohri, and M. Saraclar, "General indexation of weighted automata: application to spoken utterance retrieval," in *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL*, 2004.

[10] C. Liu, D. Wang, and J. Tejedor, "N-gram FST indexing for spoken term detection," in *Interspeech*, 2012.

[11] B. Logan, P. Moreno, J.-M. Van Thong, and E. Whittaker, "An experimental study of an audio indexing system for the web," in *ICSLP*, 2000.

[12] M. Akbacak, D. Vergyri, and A. Stolcke, "Open-vocabulary spoken term detection using graphone-based hybrid recognition systems," in *ICASSP*, 2008.

[13] Y.-C. Pan, H.-L. Chang, and L.-S. Lee, "Subword-based position specific posterior lattices (S-PSPL) for indexing speech information," in *Interspeech*, 2007.

[14] B. Logan, J.-M. Van Thong, and P. Moreno, "Approaches to reduce the effects of OOV queries on indexed spoken audio," *Multimedia, IEEE Transactions on*, vol. 7, pp. 899 – 906, 2005.

[15] R. Wallace, R. Vogt, and S. Sridharan, "A phonetic search approach to the 2006 NIST spoken term detection evaluation," in *Interspeech*, 2007.

[16] V. T. Turunen, "Reducing the effect of OOV query words by using morph-based spoken document retrieval," in *Interspeech*, 2008.

[17] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *SIGIR*, 2007.

[18] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *ICASSP*, 2008.

[19] Y. Itoh, K. Iwata, K. Kojima, M. Ishigame, K. Tanaka, and S.-W. Lee, "An integration method of retrieval results using plural subword models for vocabulary-free spoken document retrieval," in *Interspeech*, 2007.

[20] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *ICASSP*, 2006.

[21] K. Ng, "Subword-based approaches for spoken document retrieval," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[22] S.-W. Lee, K. Tanaka, and Y. Itoh, "Combining multiple subword representations for open-vocabulary spoken document retrieval," in *ICASSP*, 2005.

[23] S. Meng, P. Yu, J. Liu, and F. Seide, "Fusing multiple systems into a compact lattice index for Chinese spoken term detection," in *ICASSP*, 2008.

[24] Y.-C. Pan, H.-Y. Lee, and L.-S. Lee, "Interactive spoken document retrieval with suggested key terms ranked by a Markov decision process," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 632 –645, 2012.

[25] S. Parlak and M. Saraclar, "Spoken information retrieval for Turkish broadcast news," in *SIGIR*, 2009.

[26] S.-Y. Kong, M.-R. Wu, C.-K. Lin, Y.-S. Fu, and L.-S. Lee, "Learning on demand - course lecture distillation by information extraction," in *ICASSP*, 2009.

[27] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *Interspeech*, 2007.

[28] J. H. L. Hansen, R. Huang, P. Mangalath, B. Zhou, M. Seadle, and J. R. Deller, "SPEECHFIND: Spoken document retrieval for a national gallery of the spoken word," 2004.

[29] M. Goto, J. Ogata, and K. Eto, "Podcastle: A web 2.0 approach to speech recognition research," in *Interspeech*, 2007.

[30] C. Alberti, M. Bacchiani, A. Bezman, C. Chelba, A. Drofa, H. Liao, P. Moreno, T. Power, A. Sahuguet, M. Shugrina, and O. Siohan, "An audio indexing system for election video material," in *ICASSP*, 2009.

[31] H.-Y. Lee, T.-H. Wen, and L.-S. Lee, "Improved semantic retrieval of spoken content by language models enhanced with acoustic similarity graph," in *SLT*, 2012.

[32] T.-W. Tu, H.-Y. Lee, Y.-Y. Chou, and L.-S. Lee, "Semantic query expansion and context-based discriminative term modeling for spoken document retrieval," in *ICASSP*, 2012.

[33] H.-L. Chang, Y.-C. Pan, and L.-S. Lee, "Latent semantic retrieval of spoken documents over position specific posterior lattices," in *SLT*, 2008.

[34] 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, pp. 2602 – 2612, 2012.

[35] X. Hu, R. Isotani, H. Kawai, and S. Nakamura, "Cluster-based language model for spoken document retrieval using NMF-based document clustering," in *Interspeech*, 2010.

[36] T. Akiba and K. Honda, "Effects of query expansion for spoken document passage retrieval," in *Interspeech*, 2011.

[37] R. Masumura, S. Hahm, and A. Ito, "Language model expansion using webdata for spoken document retrieval," in *Interspeech*, 2011.

[38] H. Nishizaki, K. Sugimotoy, and Y. Sekiguchi, "Web page collection using automatic document segmentation for spoken document retrieval," in *APSIPA*, 2011.

[39] 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.

[40] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *SIGIR*, 2006.

[41] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing," in *SIGIR*, 2011.

[42] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in *SIGIR*, 2007.

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

[44] V. Lavrenko and W. B. Croft, "Relevance-based language models," in *SIGIR*, 2001.

[45] Y. Lv and C. 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*, ser. CIKM '09, 2009.

[46] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "Statistical lattice-based spoken document retrieval," *ACM Trans. Inf. Syst.*, vol. 28, pp. 2:1–2:30, 2010.

[47] H.-Y. Lee, P.-W. Chou, and L.-S. Lee, "Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity," in *Interspeech*, 2012.

[48] H.-Y. Lee, Y.-N. Chen, and L.-S. Lee, "Improved speech summarization and spoken term detection with graphical analysis of utterance similarities," in *APSIPA*, 2011.

[49] Y.-N. Chen, C.-P. Chen, H.-Y. Lee, C.-A. Chan, and L.-S. Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *ICASSP*, 2011.

[50] C. Zhai, "Statistical language models for information retrieval a critical review," *Found. Trends Inf. Retr.*, vol. 2, pp. 137–213, 2008.

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

[52] D. Karakos, M. Dredze, K. Church, A. Jansen, and S. Khudanpur, "Estimating document frequencies in a speech corpus," in *ASRU*, 2011.

[53] H.-Y. Lee, C.-P. Chen, C.-F. Yeh, and L.-S. Lee, "Improved spoken term detection by discriminative training of acoustic models based on user relevance feedback," in *Interspeech*, 2010.

[54] H.-Y. Lee, C.-P. Chen, and L.-S. 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.

[55] A. Norouzian, A. Jansen, R. Rose, and S. Thomas, "Exploiting discriminative point process models for spoken term detection," in *Interspeech*, 2012.

[56] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, pp. 135–161, 2005.

[57] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. of Uncertainty in Artificial Intelligence*, 1999.

[58] A. Atreya and C. Elkan, "Latent semantic indexing (LSI) fails for TREC collections," *SIGKDD Explor. Newsl.*, vol. 12, pp. 5–10, 2011.

[59] H. M. Wang, B. Chen, J. W. Kuo, and S. S. Cheng, "MATBN: a Mandarin Chinese broadcast news corpus," in *Comput. Linguistics Chinese Language Process.*, 2005, pp. 219 – 236.

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

[61] *http://www.lemurproject.org/lemur/retrieval.php*.

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

[63] ——, "Model-based unsupervised spoken term detection with spoken queries," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 1330–1342, 2013.

[64] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments," in *Information Processing and Management*, 2000.

[65] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, ch. 12, pp. 248 – 250.

Hung-yi Lee, was born in 1986. He received the M.S. and Ph.D. degrees in communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2010 and 2012, respectively. From September 2012 to August 2013, he was a postdoctoral fellow in Research Center for Information Technology Innovation, Academia Sinica. He is currently visiting the Spoken Language Systems Group of MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). His research focuses on spoken content retrieval and spoken document summarization.

Lin-shan Lee, (F3) received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA.

He has been a Professor of electrical engineering and computer science at the National Taiwan University, Taipei, Taiwan, since 1982 and holds a joint appointment as a Research Fellow of Academia Sinica, Taipei. His research interests include digital communications and spoken language processing. He developed several of the earliest versions of Chinese spoken language processing systems in the world including text-to-speech systems, natural language analyzers, dictation systems, and voice information retrieval systems.

Dr. Lee was Vice President for International Affairs (1996-1997) and the Awards Committee chair (1998-1999) of the IEEE Communications Society. He was a member of the Board of International Speech Communication Association (ISCA 2002-2009), a Distinguished Lecture (2007-2008) and a member of the Overview Paper Editorial Board (since 2009) of the IEEE Signal Processing Society, and the general chair of ICASSP 2009 in Taipei. He is a fellow of ISCA since 2010, and received the Meritorious Service Award from IEEE Signal Processing Society in 2011.