Improved Speech Summarization and Spoken Term Detection with Graphical Analysis of Utterance Similarities

Hung-yi Lee, Yun-nung Chen, and Lin-shan Lee
National Taiwan University
E-mail: {tlkagkb93901106, vivian.ynchen}@gmail.com, lslee@gate.sinica.edu.tw

***Abstract*—We present summarization and spoken term detection (STD) approaches that take into account similarities between utterances to be scored for summary extraction or ranking in STD. A graph is constructed in which each utterance is a node. Similar utterances are connected by edges, with the edge weights representing the degree of similarity. The similarity for summarization is topical similarity; that for STD is feature-space similarity. The score of each utterance for extraction in summarization and ranking in STD is not solely decided by the individual utterance but is influenced by similar utterances on the graph. Experimental results show significant improvements compared with two baselines in terms of the ROUGE evaluation for summarization and mean average precision for STD.**

I. Introduction

In the Internet era, digital network content covers all of the information and activities in human life. The most attractive form of network content is multimedia, including speech. The subjects, topics, and core concepts of such speech information is usually to be found within the content itself. However, as multimedia or spoken documents are just video or audio signals, they are usually much more difficult to retrieve and browse, because they cannot be easily displayed on-screen, and the user cannot simply "skim through" each one from beginning to end. Hence the importance of speech information retrieval and spoken document summarization in helping users efficiently mine speech content [1].

In general, there are two stages in a speech information retrieval system [2]. In the first stage, the audio content is recognized and transformed into transcriptions or lattices. 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 utterances to the user. The discussion in this paper is limited to spoken term detection (STD) [3], in which the query is a term submitted by the user in text form and the system returns a list of spoken utterances containing that term. Summarization is the process of automatically creating a compressed version of a given spoken document that provides useful information for the user. The information content of a summary depends on the user's needs. Here we discuss topic-oriented summaries; we thus focus on extracting the information in the document that is related to the specified topic. Like speech information retrieval, the spoken documents are first transcribed into text using the recognition engine, and then the system selects a number of indicative utterances from the original spoken documents according to a target summarization ratio, and contatenates them to form a summary.

Both STD and summarization can be considered utterance ranking problems which rank the utterances based on cues found in the utterance set but with different ranking targets. In STD, the system ranks the utterances over the entire spoken archive based on the relevance scores assigned to the utterances representing the probabilities of the appearance of the query term. The posterior probability of the query term derived from the lattice is widely used as a relevance score [4], [5]; other confidence measures are also useful [6]. In summarization, each utterance is given an importance score representing how well it represents the document as a whole. The utterances are selected for the summary in the order of the ranking of the importance scores until the number of terms in the summary exceeds a target summarization ratio. The importance score of each utterance is typically based on its grammatical structure as well as various statistical measures, linguistic measures, and confidence scores of the terms in the utterance [7]. In mainstream approaches for STD and summarization, the utterance ranking score relies only on evidence observed in each individual utterance; we however believe that the relationships among utterances may yield fruitful information for ranking and therefore should not be ignored.

Although much research has been devoted to ranking instances, additionally taking into account such inter-instance

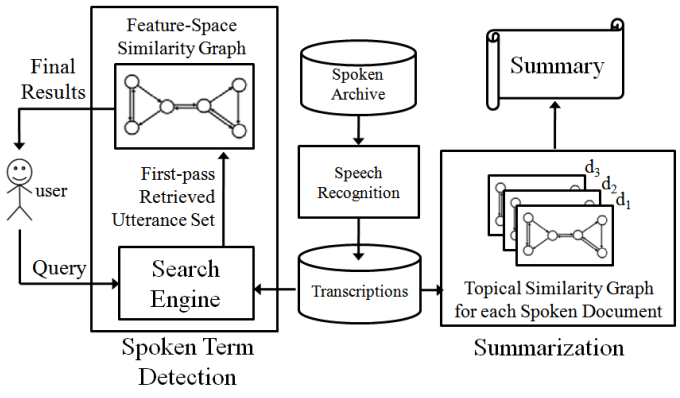

Fig. 1. *The framework for the proposed approach.*

relationships significantly complicates matters. These relationships are usually represented as a graph in which each node is an instance; ranks are induced by the scores assigned to the nodes. The relationships between these instances are represented by edges, the weights of which reflect the degree of relation. "Similarity" between instances is especially useful for ranking. We assume that similar instances should be ranked similarly because they share the same properties. Thus we assign high scores to instances that are connected to other instances with high scores; whereas instances that are similar to other instances with low scores are assigned low scores. In other words, each instance's score is influenced by other instances that share similar properties. Although similarities between instances are usually helpful in ranking problems, what kinds of similarities as edges are able to improve the ranking performance should be individually researched for each domain.

We attempt to take into account similarity between utterances in summarization and STD. In topic-oriented summarization, where the summary is expected to describe specific information about the topic of the document, the utterances in the summary should have similar content and thus similar latent topic distributions. Hence we seek to extract utterances with similar latent topics as the summary. Therefore, here the similarity between utterances being considered refers to topical similarity in summarization, and an utterance with strong evidences to be selected to form a summary would increase the summary selection possibility of the utterances with similar latent topics. For spoken term detection, it has been verified by the feature-space pseudo-relevance feedback (PRF) techniques proposed earlier [8] that utterances that are highly similar at the feature level to parts of the pseudo-relevant utterances are likely to be relevant. Hence, for STD, similarity between utterances refers to feature-space similarity of query hypotheses. The system assigns the utterances close relevance scores if they are characterized by similar features.

We present a graph-based method to take the advantage of utterance similarity in STD [9] and summarization [10]. The frameworks of the proposed approaches for STD and summarization are shown in Fig. 1, with summarization on the right. An utterance graph is constructed for each transcribed spoken document. The nodes on the graph are utterances in one document, and topically similar utterances are connected with edges. Each utterance's importance score depends not only on the statistical measure of the terms in the utterance itself but also on utterances connected to it within the graph. For STD, on the left in Fig. 1, when the user enters a query, the retrieval engine searches through the lattices to find the utterances containing the query term as the first-pass returned list, which is not shown to the user. A graph is constructed from this list in which each node represents an utterance in the list and the edges represent high feature-space similarity between utterances. The relevance score of each utterance is partially decided by the utterances connected, and then the list is reranked accordingly.

The proposed approaches for summarization and STD are in Section II and Section III respectively. In Section IV, course lectures were taken as an example in the experiments to test the proposed approach for both STD and summarization. In Section V we offer concluding remarks.

II. Summarization with Utterance Similarity

We introduce a graph-based method that takes into account topical similarity when computing importance scores for the utterances in a spoken document. Whereas similar approaches to utterance similarity have been used on text summarization [11], [12], previous research on text only used term similarity instead of the proposed latent topic similarity.

A. Baseline

Here the statistical measures of the terms in utterance x belonging to spoken document d are used to infer importance score $I_d(x)$:

$$I_d(x) = \sum_{t_i \in x} n(t_i, x) s(t_i, d), \tag{1}$$

where $n(t_i, x)$ is the occurrence count of term t_i in utterance x, and $s(t_i, d)$ is the statistical measure of term t_i. In this work, $s(t_i, d)$ is defined as described below in two different ways.

1) Latent Topic Entropy-based Statistical Measure: Probabilistic latent semantic analysis (PLSA) [13] has been widely used to analyse the semantics of documents based on a set of latent topics. Given a set of documents $\{d_j, j = 1, 2, ..., J\}$ and all the terms $\{t_i, i = 1, 2, ..., M\}$ they include, PLSA uses a set of latent topic variables, $\{T_k, k = 1, 2, ..., K\}$, to characterize the "term-document" co-occurrence relationships. The probability of observing term t_i given document d_j can be parameterized by

$$P(t_i|d_j) = \sum_{k=1}^{K} P(t_i|T_k) P(T_k|d_j). \tag{2}$$

The PLSA model can be optimized using the EM algorithm by maximizing a likelihood function.

The latent topic entropy (LTE), $LTE(t_i)$, for a given term t_i can be calculated as (3) from the topic distribution $P(T_k|t_i)$ for each term t_i:

$$LTE(t_i) = -\sum_{k=1}^{K} P(T_k|t_i) \log P(T_k|t_i), \tag{3}$$

where the topic distribution $P(T_k|t_i)$ can be estimated as follows [14], [15]:

$$P(T_k|t_i) = \frac{P(t_i|T_k) \times P(T_k)}{P(t_i)} \simeq \frac{P(t_i|T_k)}{P(t_i)}, \tag{4}$$

where the probability $P(T_k)$ is omitted because there is as yet no good approach to estimate it. $P(t_i)$ can be obtained from a large corpus. $LTE(t_i)$ is a measure of how focused the term t_i is on a few topics; a lower latent topic entropy implies the term carries more topical information.

The statistical measure $s(t_i, d)$ in (1) can be defined based on $LTE(t_i)$ in (3) as

$$s_{LTE}(t_i, d) = \frac{n(t_i, d)}{LTE(t_i)}, \tag{5}$$

where $n(t_i, d)$ is the occurrence count of term t_i in document d. Score $s_{LTE}(t_i, d)$ is inversely proportional to $LTE(t_i)$. Previous work [14], [15] has showed that this measure outperforms the very successful "significance score" in speech summarization [7]. Here we use $s_{LTE}(t_i, d)$ as one of the baselines.

2) Key-Term-based Statistical Measure: Key terms are the terms in a document that carry the core concepts of the content. They are useful for indexing, retrieval, and browsing. In general there are two types of key terms: keywords (single words) and key phrases (such as "hidden Markov model"). Automatically extracting key terms from spoken content is still a difficult problem, but some initial approaches have been shown to be successful in recent experiments [16]. Such approaches include the use of right/left branching entropy derived from PAT-Trees to extract frequently occurring patterns including two or more words, and identifying or verifying key terms (including key phrases) by prosodic (pitch, duration, energy), lexical, and semantic (from PLSA) features with unsupervised techniques or supervised training. Such automatically extracted key terms are very helpful in summarization.

With key terms thus automatically extracted (with some errors), we can estimate a new latent topic probability $P_{KEY}(T_k|d)$ that is hopefully better than the $P(T_k|d)$ calculated directed from the PLSA model:

$$P_{KEY}(T_k|d) = \frac{\sum_{t_i \in key} n(t_i, d) P(T_k|t_i)}{\sum_{k=1}^{K} \sum_{t_i \in key} n(t_i, d) P(T_k|t_i)}, \tag{6}$$

where key is the set of automatically extracted key terms, and $P(T_k|t_i)$ is in (4). Therefore only the automatically extracted key terms t_i in d are considered, eliminating the influence from other insignificant terms. We then define the statistical measure $s(t_i, d)$ as

$$s_{KEY}(t_i, d) = \sum_{k=1}^{K} LTS_{t_i}(T_k) P_{KEY}(T_k|d). \tag{7}$$

$LTS_{t_i}(T_k)$, the latent topic significance (LTS) for term t_i with respect to topic T_k, is defined [14], [15] as

$$LTS_{t_i}(T_k) = \frac{\sum_{d_j \in D} n(t_i, d_j) P(T_k|d_j)}{\sum_{d_j \in D} n(t_i, d_j)[1 - P(T_k|d_j)]}, \tag{8}$$

where $n(t_i, d_j)$ is the occurrence count of term t_i in document d_j. In the numerator of (8), the count of term t_i in document d_j, $n(t_i, d_j)$, is weighted by the likelihood that topic T_k is addressed by document d_j, $P(T_k|d_j)$, and then summed over all documents d_j in the PLSA model training corpus $\mathcal{D}$. Therefore the numerator is the total count of term t_i used for the given topic T_k over the whole PLSA training corpus, as estimated by PLSA model. The denominator is very similar except that it is for latent topics other than T_k, so $P(T_k|d_j)$ is replaced with $[1 - P(T_k|d_j)]$. Thus, a higher $LTS_{t_i}(T_k)$ indicates the term t_i is more significant for latent topic T_k.

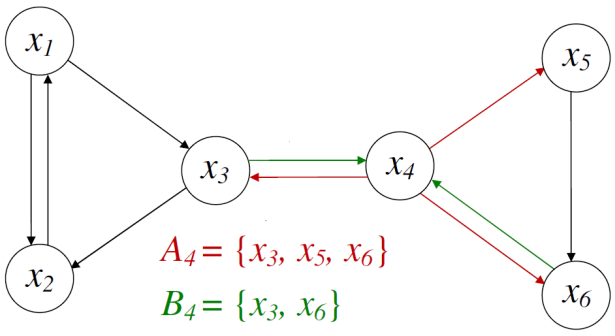

Fig. 2. *A simplified example of a graph, the nodes of which correspond to utterances. A_i and B_i are the node sets connected respectively by outgoing and incoming edges of x_i.*

B. Proposed Approach

To take into account the topical similarity between utterances during summarization, a graph representing the topical similarity between utterances in document d is first constructed. All the utterances in d are nodes on the graph, and each utterance connects to only the N_1 most topically similar utterances. Then an importance score is assigned to each utterance based on the graph structure. The weights of the edges correspond to the topical similarity between the associated utterances. To estimate the topical similarity between two utterances, we first compute the probability that topic T_k is addressed by utterance x_i,

$$P(T_k|x_i) = \frac{\sum_{t \in x_i} n(t, x_i) P(T_k|t)}{\sum_{t \in x_i} n(t, x_i)}. \tag{9}$$

Then the edge weight for utterance x_i to x_j (with direction $x_i \rightarrow x_j$) is defined by accumulating $LTS_t(T_k)$ in (8) weighted by $P(T_k|x_i)$ for all terms t in x_j over all latent topics,

$$W_{topic}(x_i, x_j) = \sum_{t \in x_j} \sum_{k=1}^{K} LTS_t(T_k) P(T_k|x_i). \tag{10}$$

A simplified example for such graph is given in Figure 2, in which A_i and B_i are the utterance sets connected by outgoing and incoming edges of utterance x_i respectively.

Consider document d with utterances $\{x_i, i = 1, 2, ..., N_d\}$. The proposed importance scores are $\{I_d^G(x_i), i = 1, 2, ..., N_d\}$ satisfying the equation

$$I_d^G(x_i) = (1 - \alpha) \hat{I}_d(x_i) + \alpha \sum_{x_j \in B_i} P_{topic}(j, i) I_d^G(x_j) \tag{11}$$

for $i = 1, 2, ..., N_d$. B_i is the set of utterances connected to utterance x_i via incoming edges. $\hat{I}_d(x_i)$ is the normalized importance score

$$\hat{I}_d(x_i) = \frac{I_d(x_i)}{\sum_{j=1}^{N_d} I_d(x_j)}. \tag{12}$$

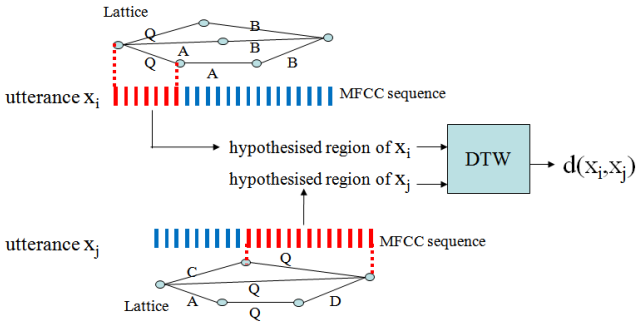

Fig. 3. *The definition of "hypothesised region"(the red part) of an utterance x_i and the distance $d(x_i, x_j)$ between two utterances x_i and x_j. The hypothesised region of an utterance x_i is the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query Q with the highest posterior probability in the lattice.*

$I_d(x_j)$ can use either a LTE-based or key-term-based statistical measure. $P_{topic}(j,i)$ is the weight of the edge from x_j to x_i normalized by the weights over the outgoing edges of utterance x_j:

$$P_{topic}(j,i) = \frac{W_{topic}(x_j, x_i)}{\sum_{x_k \in A_j} W_{topic}(x_j, x_i)}, \tag{13}$$

where A_j are the utterances connected by the outgoing edges of x_k. α is an interpolation weight between 0 and 1. Thus $I_d^G(x_i)$, the new importance score of x_i, takes into account not only the statistical measures of the terms in x_i but also the importance scores of the utterances that are topically very similar to x_i. The higher the edge weight, that is, the topical similarity, the more influence it has on the importance score of x_i. During summarization, with the graph-based (11), all utterances in the document d are taken into consideration jointly and not individually. The normalizations in (12) and (13) are necessary to formulate (11) as the random walk problem [17], [18]. The theory of the random walk guarantees that $\{I_d^G(x_i), i = 1, 2, ..., N_d\}$ satisfying (11) is unique and nonnegative, which can be found efficiently by the power method [19].

For better results, $I_d^G(x_i)$ is integrated with the baseline $I_d(x_i)$ as

$$I'_d(x_i) = I_d(x)(I_d^G(x_i))^{\delta_1}, \tag{14}$$

where δ_1 is a weighting parameter. The proposed approach uses $I'_d(x_i)$ as the importance score when ranking the utterances in a spoken archive for summary selection.

III. Spoken Term Detection with Utterance Similarity

When the user submits query Q, the retrieval engine searches over all of the lattices to find those utterances containing the query Q as the first-pass returned list ranked by the relevance score $S_Q(x)$. The spoken segment set retrieved in the first pass is denoted as X_Q. With the success of acoustic feature space pseudo-relevance feedback (PRF) [8], we know that if an utterance has a "hypothesized region" very "similar" to utterances with high relevance scores, it is more likely to be relevant, or its relevance score should be increased. A hypothesized region (Fig. III) is the most probable occurrence of query Q in the utterance, as the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query term Q with the highest posterior probability in the lattice. As shown in Fig. III, the distance $d(x_i, x_j)$ between two utterances x_i and x_j given query Q is the dynamic time warping distance [20] of the MFCC sequences corresponding to the time spans of the hypothesized region. The feature-space similarity between the utterances x_i and x_j is defined according to $d(x_i, x_j)$. Here we introduce a graph-based approach for reranking the first-pass returned list with feature-space similarity.

A. Baseline

We use as the first baseline the first-pass retrieval result, ranked according to the widely-used query posterior probability. PRF is used as the second set of baselines.

1) First Pass: The relevance score $S_Q(x)$ of utterance x with respect to query Q is defined as

$$S_Q(x) = \sum_{word(a)=Q} P(a|x), \tag{15}$$

where a is any arc in the lattice of x, $word(a)$ is the word hypothesis of a and $P(a|x)$ is the posterior probability. The first pass retrieval result is ranked according to $S_Q(x)$.

2) Pseudo-Relevance Feedback: In PRF, a pseudo-relevant utterance set Y_Q is selected out of the first-pass retrieval result X_Q for query Q, and the similarity between each utterance x_i in the first-pass returned list and the set Y_Q is computed and integrated with the original relevance score. The top N utterances in the returned list (that is, the N utterances with highest relevance scores) in X_Q are selected as Y_Q. The distance between utterance x_i and set Y_Q is

$$D(x_i, Y_Q) = \sum_{x_j \in Y_Q} d(x_i, x_j)^2, \tag{16}$$

the total distance between x_i and all utterances in Y_Q. The value of $D(x_i, Y_Q)$ is normalized between 0 and 1 as $\hat{D}(x_i, Y_Q)$, and the similarity between x_i and Y_Q is

$$SIM(x_i, Y_Q) = 1 - \hat{D}(x_i, Y_Q), \tag{17}$$

which is 1 minus the normalized distance between x_i and Y_Q. The retrieval result is ranked according to $S_Q(x)SIM(x_i, Y_Q)^{\delta}$, where $S_Q(x)$ is defined in (15).

B. Proposed Approach

To take into account utterance similarities in STD ranking, we first construct a graph representing the feature-space similarities between the utterances of the first-pass retrieved utterances X_Q. Each utterance in X_Q is a node in the graph, and each utterance (node) connects to the most similar N_2 utterances in feature space. The weight of the edge from utterance x_i to x_j is

$$W_{sim}(x_i, x_j) = 1 - \frac{d(x_i, x_j) - d_{min}}{d_{max} - d_{min}}, \tag{18}$$

where d_{max} and d_{min} are the largest and smallest values of $d(x_i, x_j)$ for all pairs of utterances in X_Q. [1] Then new feature-space similarity-based relevance scores are obtained based on the graph structure. This graph is the same as Fig. 2 but uses edges that instead represent feature-space similarities. The definitions of A_i and B_i are the same as in Section II-B.

The proposed relevance scores $\{R_Q^G(x_i), x_i \in X_Q\}$ compose the value set satisfying

$$R_Q^G(x_i) = (1-\alpha)\hat{R}_Q(x_i) + \alpha \sum_{x_j \in B_i} P_{sim}(j,i) R_Q^G(x_j) \quad (19)$$

for all $x_i \in X_Q$.

$$\hat{R}_Q(x_i) = \frac{S_Q(x_i)}{\sum_{x_j \in X_Q} S_Q(x_j)} \quad (20)$$

is the normalized relevance score of utterance x_i, $S_Q(x)$ is as defined in (15), and X_Q is the first-pass retrieved utterance set. $P_{sim}(j,i)$ is the normalization of the edge weight $W_{sim}(x_j, x_i)$ over the outgoing edges of utterance x_j on the graph:

$$P_{sim}(j,i) = \frac{W_{sim}(x_j, x_i)}{\sum_{x_k \in A_j} W_{sim}(x_j, x_k)}, \quad (21)$$

where A_j are again the utterances connected by the outgoing edges of x_k. α is an interpolation weight between 0 and 1. Equation (19) shows that $R_Q^G(x_i)$, the new relevance score of utterance x_i, depends on two factors. One is the posterior probability of query Q in the x_i lattice (the first term on the right side of (19)), and the other is the relevance scores of the similar utterances (the second term on the right side). Compared with PRF, which takes into account only similarities to utterances in the pseudo-relevant set, the proposed approach takes into consideration the relations of all the utterances retrieved. To be specific, PRF only raises the scores of utterances that are connected to other utterances with high relevance scores; the proposed method, however, also lowers the relevance scores of those utterances that are connected to other utterances with low relevance utterances. Therefore, the proposed approach outperforms PRF. The normalization in (21) and (20) here formulates (19) as a random walk problem on a graph. As mentioned in Section II-B, the solution of $\{R_Q^G(x_i), x_i \in X_Q\}$ satisfying (19) is unique and nonnegative.

$R_Q^G(x_i)$ is integrated with the original relevance score $S_Q(x_i)$ for re-ranking as

$$S'_Q(x_i) = S_Q(x_i)(R_Q^G(x_i))^{\delta_2}, \quad (22)$$

where δ_2 is a weighting parameter. The final retrieval result displayed to the user is then ranked according $S'_Q(x_i)$.

IV. Experiments

A. Corpus

As the testing archive we used a corpus of 33 hours of recorded lectures for a course offered at National Taiwan University produced by a single instructor. Used for retrieval and summarization, this corpus is quite noisy and spontaneous. The lectures were given in Mandarin Chinese (the "host" language) with English (the "embedded" language) terms and phrases embedded within the Mandarin utterances.

B. Summarization

1) Experimental Setup: For summary experiments, both manual transcriptions (Manual) without word errors and the results of speech recognition (ASR) on the lectures were used for testing. For speech recognition, the acoustic model was trained using the maximum likelihood criterion with 4602 state-tied triphones spanned from 37 monophones using a corpus of noiseless Mandarin read speech, including 24.6 hours of data produced by 100 males and 100 females, and adapted with a 25.2-minutes bilingual corpus from the target speaker (the course instructor) [21]. The language model was trained with two other courses offered by the same instructor and was adapted to the course slides. The accuracies for the ASR transcriptions were 78.15% for Mandarin characters, 53.44% for English words, and 76.26% overall. The unsupervised automatic key term extraction approach mentioned in Section II-A2 was used, and both keywords and key phrases were extracted [16]. The key terms F-measures for ASR and manual transcriptions were 52.60% and 55.84% respectively.

To evaluate the performance of the automatically generated summaries, we used the well-known evaluation package ROUGE [22]. The ROUGE-N F-measure ($N = 1, 2, 3$) and ROUGE-L were used to evaluate summarization results. We segmented the whole lecture into 155 documents using topic segmentation [23], and extracted the summary for each document. As the test corpus we used 34 out of the 155 documents for which reference summaries were produced manually. We used 32 topics for PLSA, and set α to 0.9. In the experiments presented below, the summarization ratio was set to 10%, 20%, and 30% respectively. The automatically extracted key phrases (all with more than one word) were taken as individual terms in PLSA modeling and all following processes.

2) Experimental Results: Fig. 4 shows the results for ROUGE-N and ROUGE-L for ASR ((a)–(d)) and manual transcriptions ((e)–(h)). In each case the three groups of bars are for 10%, 20%, and 30% summarization ratios, and in each group the four bars are respectively for the LTE-based statistical measure $s_{LTE}(t_i, d)$ in (5), that followed by the proposed topical similarity graph (LTE + G), the key-term-based statistical measure in (7) (Key) and that followed by the proposed approach (Key + G). In all cases, the key-term-based statistical measure (bar 3) outperformed the LTE baseline (bar 1). Clearly key term knowledge was very helpful, especially for manual transcriptions. This is probably because in manual transcriptions all key terms were correctly transcribed (although they were sometimes incorrectly extracted), which ensured more accurate estimation of the key-term-based statistical measures. Similar but slightly less significant improvements were yielded for ASR transcriptions.

In all cases, the proposed approach considering utterance relationships based on topical similarity graph improved on

[1] $W_{sim}(x_i, x_j)$ and $W_{sim}(x_j, x_i)$ are equal.

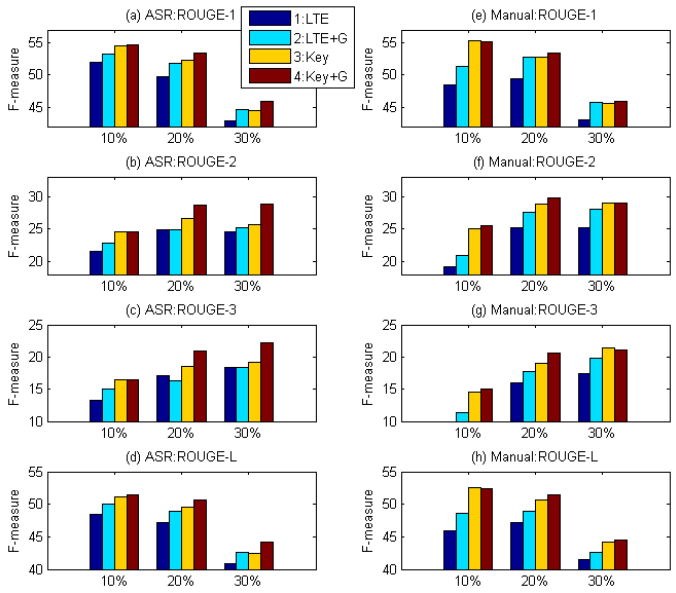

Fig. 4. *The results of different choices of parameters: LTE-based (1, 2) or key-term-based (3, 4), with (2, 4) or without (1, 3) the topical similarity graph, for ASR ((a)–(d)) or manual ((e)–(h)) transcriptions at summarization ratios of* 10%, 20%, *and* 30%.

the LTE-based statistical measure (bar 2 vs bar 1), except for ASR ROUGE-3 with the 20% summarization ratio. For ASR transcriptions ((a)–(d)), the proposed approach also improved on the key-term-based statistical measure (bar 4 vs bar 3). However, at the 10% and 30% summarization ratios for manual transcriptions ((e)–(h)), the proposed approach did not similarly with help the key-term-based statistical measure (bar 4 vs bar 3). This may be because for manual transcriptions, important utterances were already well represented by the key-term-based statistical measure; hence adding extra topical similarity among utterances did not lead to better performance.

C. Spoken Term Detection

1) Experimental Setup: A tri-gram language model trained on news data was used in speech recognition. In order to evaluate the performance of the proposed approach with respect to acoustic models of different matched conditions, we used three sets of acoustic models:

- A speaker-independent model (SI) trained on 24.6 hours of read speech produced by 100 male and 100 female speakers.
- An MLLR model (MLLR) adapted from the above SI model on 500 utterances taken from the training set of the lecture corpus.
- A speaker-dependent model (SD) trained on 12 hours of the training set of the lecture corpus, all produced by the same speaker as that in the retrieval corpus.

For all acoustic models, we trained 4602 state-tied triphones spanned from 37 monophones. The recognition accuracies were 50.26%, 62.55%, and 81.34% respectively for the SI, MLLR, and SD models.

TABLE I

MAP results for the baseline, PRF, and the proposed approach with various acoustic models.

Methods	SI		MLLR		SD	
	MAP	Impr.	MAP	Impr.	MAP	Impr.
First pass	45.47	-	55.54	-	73.52	-
PRF	52.10	6.63	61.59	6.05	75.78	2.26
Proposed	53.42	7.95	63.78	8.24	76.71	3.19

Mean average precision (MAP) was used as the measure for retrieval performance evaluation. 162 Mandarin queries were manually selected in the tests, each being a single word.

2) Experimental Results: The results of the first-pass retrieval for the three sets of acoustic models are listed in the first row of Table I as the first baseline. Clearly the performance is heavily dependent on the quality of the acoustic model. The second row is PRF (described in Section III-A2), which outperforms the baseline regardless of the quality of the acoustic model. PRF serves as the second baseline.

The results of integrating the original score in (15) with the proposed scores satisfying (19) are shown in the third row of Table I. These results show that the integration with the scores derived based on utterance similarity yields better performance than the first-pass results for all acoustic models, especially with poorer acoustic models (SI and MLLR), or when the original relevance scores are less precise. It also clearly outperformed the PRF approach. This shows the effectiveness of taking into account the complete relationships between all the utterances retrieved.

V. Conclusions

We present graph-based approaches that take into account utterance similarity for summarization and STD. All approaches take utterances as nodes on the graph. Edge weights for summarization are represented by topical similarities; those for STD are represented by feature-space similarities. Encouraging results were obtained in the experiments for both tasks.

References

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

[2] C. Chelba, T.J. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," in *Signal Processing Magazine, IEEE*, 2008.

[3] National Institute of Standards and Technology, *The spoken term detection (STD) 2006 evaluation plan*, 2006.

[4] Ciprian Chelba and Alex Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005.

[5] Murat Saraclar, "Lattice-based search for spoken utterance retrieval," in *In Proceedings of HLT-NAACL 2004*, 2004.

[6] Dong Wang, Javier Tejedor, Joe Frankel, Simon King, and Jose Colas, "Posterior-based confidence measures for spoken term detection," in *ICASSP*, 2009.

[7] S. Furui, T. Kikuchi, Y. Shinnaka, and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, pp. 401 – 408, 2004.

[8] Chia-Ping Chen, Hung-Yi Lee, Ching-Feng Yeh, and Lin-Shan Lee, "Improved spoken term detection by feature space pseudo-relevance feedback," in *INTERSPEECH*, 2010.

[9] Yun-Nung Chen, Chia-Ping Chen, Hung-Yi Lee, Chun-An Chan, and Lin-Shan Lee, "Improved spoken term detection with graph-based re-ranking in feature space," in *ICASSP*, 2011.
[10] Yun-Nung Chen, Yu Huang, Ching-Feng Yeh, and Lin-Shan Lee, "Spoken lecture summarization by random walk over a graph constructed with automatically extracted key terms," in *reviewing*, 2011.
[11] Günes Erkan and Dragomir R. Radev, "Lexrank: graph-based lexical centrality as salience in text summarization," *J. Artif. Int. Res.*, vol. 22, pp. 457–479, 2004.
[12] Jahna Otterbacher, Gunes Erkan, and Dragomir R. Radev, "Biased lexrank: Passage retrieval using random walks with question-based priors," *Inf. Process. Manage.*, vol. 45, pp. 42–54, 2009.
[13] Thomas Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999.
[14] Sheng-Yi Kong and Lin shan Lee, "Improved spoken document summarization using probabilistic latent semantic analysis (PLSA)," in *ICASSP*, 2006.
[15] Sheng yi Kong and Lin shan Lee, "Improved summarization of chinese spoken documents by probabilistic latent semantic analysis (PLSA) with further analysis and integrated scoring," in *SLT*, 2006.
[16] Yun-Nung Chen, Yu Huang, Sheng-Yi Kong, and Lin-Shan Lee, "Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features," in *SLT*, 2010.
[17] Laszlo Lovasz, "Random walks on graphs: A survey," 1993.
[18] Brin S. and Page L., "The anatomy of a largescale hypertextual web search engine," in *WWW*, 1998.
[19] Amy N. Langville and Carl D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, pp. 135–161, January 2005.
[20] Chun-An Chan and Lin-Shan Lee, "Unsupervised spoken term detection with spoken queries using segment-based dynamic time warping," in *InterSpeech*, 2010.
[21] Ching-Feng Yeh, Liang-Che Sun, Chao-Yu Huang, and Lin-Shan Lee, "Bilingual acoustic modeling with state mapping and three-stage adaptation for transcribing unbalanced code-mixed lectures," in *ICASSP*, 2011.
[22] Chin yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Workshop on Text Summarization Branches Out*, 2004.
[23] Shou-Chieh Hsu, "Topic segmentation on lecture corpus and its application," M.S. thesis, NTU, 2008.