Spoken Content Retrieval – Beyond Cascading Speech Recognition with Text Retrieval

Lin-shan Lee, *Fellow, IEEE*, James Glass, *Fellow, IEEE*, Hung-yi Lee, Chun-an Chan

***Abstract*—Spoken content retrieval refers to directly indexing and retrieving spoken content based on the audio rather than text descriptions. This potentially eliminates the requirement of producing text descriptions for multimedia content for indexing and retrieval purposes, and is able to precisely locate the exact time the desired information appears in the multimedia. Spoken content retrieval has been very successfully achieved with the basic approach of cascading automatic speech recognition (ASR) with text information retrieval: after the spoken content is transcribed into text or lattice format, a text retrieval engine searches over the ASR output to find desired information. This framework works well when the ASR accuracy is relatively high, but becomes less adequate when more challenging real-world scenarios are considered, since retrieval performance depends heavily on ASR accuracy.**

This challenge leads to the emergence of another approach to spoken content retrieval: to go beyond the basic framework of cascading ASR with text retrieval in order to have retrieval performances that are less dependent on ASR accuracy. This article is intended to provide a thorough overview of the concepts, principles, approaches, and achievements of major technical contributions along this line of investigation. This includes five major directions: (1) Modified ASR for Retrieval Purposes: cascading ASR with text retrieval, but the ASR is modified or optimized for spoken content retrieval purposes; (2) Exploiting the Information not present in ASR outputs: to try to utilize the information in speech signals inevitably lost when transcribed into phonemes and words; (3) Directly Matching at the Acoustic Level without ASR: for spoken queries, the signals can be directly matched at the acoustic level, rather than at the phoneme or word levels, bypassing all ASR issues; (4) Semantic Retrieval of Spoken Content: trying to retrieve spoken content that is semantically related to the query, but not necessarily including the query terms themselves; (5) Interactive Retrieval and Efficient Presentation of the Retrieved Objects: with efficient presentation of the retrieved objects, an interactive retrieval process incorporating user actions may produce better retrieval results and user experiences.

I. Introduction

Today the Internet has become an everyday part of human life. Internet content is indexed, retrieved, searched, and browsed primarily based on text, and the success of these capabilities has not only changed our lives, but generated a very successful global industry in Internet-based content and services. Although multimedia Internet content is growing rapidly, with shared videos, social media, broadcasts, etc., as of today, it still tends to be processed primarily based on the textual descriptions of the content offered by the multimedia providers.

As automatic speech recognition (ASR) technologies continue to advance, it is reasonable to believe that speech and text offerings will eventually be symmetric, since they are alternative representations of human language, in the spoken and written form, respectively, and the transformation between the two should be direct and straightforward. With this perspective, spoken content retrieval, or indexing and retrieving multimedia content from its spoken part, is an important key to easier browsing and retrieving of multimedia content in the future. In cases where the essence of the multimedia content is captured by its audio, especially for broadcast programs, lectures, meetings, etc., indexing and retrieving the content based on the spoken part not only eliminates the extra requirements of producing the text description for indexing purposes, but can precisely locate the exact time when the desired information appears in the multimedia. The basic scenario for spoken content retrieval is therefore the following: when the user enters a *query*, which can be either in textual or spoken form, the system is expected to search over the spoken content and return relevant hits, possibly including the corresponding multimedia (e.g., video).

In recent years, spoken content retrieval has achieved significant advances by primarily cascading ASR output with text information retrieval techniques [1]–[8]. With this approach, the spoken content is first converted into word sequences or lattices via ASR. In order to cope with ASR errors, lattices have been used to represent the spoken content instead of a single word sequence [9]–[13], and subword-based techniques have been used to some extent to address the out-of-vocabulary (OOV) problem [11]–[14]. For a subsequent user query (represented by lattices if spoken [8]), the text retrieval engine searches over the ASR output, and returns the relevant spoken content.

The cascade approach was very successful for the task of Spoken Document Retrieval (SDR, the term frequently used for this task earlier) track of Text REtrieval Conference (TREC), and achieved similar retrieval performance when compared with retrieval performance from human transcriptions. For this task, the word error rates (WERs) were 15-20%, which were not too far from the accuracy of the approximate manual transcriptions, and both the queries and target documents were relatively long which made good retrieval performance easier. Therefore, initially, spoken document retrieval was considered to be a "solved" problem [15]. Many successful applications were developed based on this framework, such as SpeechFind [16], PodCastle [17], GAudi (short for Google Audio Indexing) [18], MIT Lecture Browser [19] and NTU Virtual Instructor [20], [21]. However, the cascade approach was subsequently found to work well mainly for relatively high ASR accuracies, because the achievable retrieval performance is inevitably highly dependent on ASR quality. It naturally becomes less adequate when more challenging real-world tasks were considered, such as the use of short queries to retrieve short voice segments from telephone conversations, meetings, academic lectures, or shared videos, for spontaneous speech with OOV words, varying acoustic conditions and higher WERs [9].

One obvious solution to rectify these issues is to reduce the WERs of ASR systems. Much research continues to be devoted to reducing ASR WERs, and significant improvements continue to be achieved [22]–[37], based on which very good improvements on spoken content retrieval performance were also reported [38]–[47]. However, we must assume that spoken content on the Internet is produced by millions of different speakers, in different parts of the world, in thousands of different languages, on unlimited topics, and under widely varying acoustic conditions. It is therefore difficult to imagine that ASR technology will be capable of transcribing all such spoken content with low enough WERs to enable good spoken content retrieval. This is the motivation for the emergence of other approaches in this area as explained below.

In recent years, researchers have begun to explore alternative strategies to surmount the limitation of spoken content retrieval performance imposed by the inevitable and uncontrollable ASR errors, i.e., to find new rationales or frameworks for spoken content retrieval beyond the conventional framework of directly cascading a text retrieval engine on top of an ASR module. Several innovative directions have been developed, achieving retrieval performance less constrained by ASR accuracies. These emerging research directions are what is being referred to as "Beyond Cascading Speech Recognition with Text Retrieval" in this overview article.

This article is thus intended to provide a thorough overview of the concepts, principles, approaches, and achievements of major technical contributions along these new directions, with the hope that researchers can find it easier to explore additional possibilities for future development in the promising area of spoken content retrieval. Since this article is not going to cover all aspects of spoken content retrieval, particularly the most common approach of cascading speech recognition with text retrieval, the reader is referred to several excellent tutorial chapters and papers [1]–[3], [7]. Instead, this article will focus on the distinct subject of "Beyond Cascading Speech Recognition with Text Retrieval". This will be categorized into five major directions as very briefly summarized below.

1) Modified Speech Recognition for Retrieval Purposes: This approach uses cascading ASR and text retrieval, but the ASR module is optimized for retrieval performance. This idea originated from the observation that retrieval performance is not always directly related to ASR accuracy, which led to research aimed at jointly optimizing ASR and retrieval, instead of doing them separately.
2) Exploiting Information not present in ASR outputs: Some potentially useful information, such as the temporal structure of the signal, is inevitably lost when speech signals are decoded into phonemes or HMM states in standard ASR. Therefore, it is possible to augment ASR output with complementary information to enhance retrieval performance.
3) Directly Matching on Acoustic Level without ASR: When the query is spoken, it can be directly matched with spoken content at the acoustic level, instead of at a symbolic level, so that no standard ASR module is needed. All the problems with ASR such as recognition errors, the OOV problem, the need for matched annotated corpora for training acoustic models, etc. are all automatically eliminated.
4) Semantic Retrieval of Spoken Content: Spoken content semantically related to the query does not always contain the query terms. For example, with the query of "White House" many target objects regarding the president of United States may not include the query terms "White House" but should be retrieved. Many semantic retrieval techniques originally developed for text retrieval for such purposes are useful, but very interesting approaches specifically for spoken content were also developed.
5) Interactive Retrieval and Efficient Presentation of Retrieved Objects: The high degree of uncertainty in ASR may be properly taken care of by efficient user interaction learned from spoken dialogues. However, spoken content is difficult to display visually on a screen, and is not as easy to, for example, scan and select by a user, as compared to text. Thus, technologies such as automatic key term extraction, title generation, summarization, and semantic structuring of spoken content are crucial for user-friendly interfaces that enable easier access to the retrieved objects.

The remainder of this article is organized as follows. In Section II, we first provide some necessary background knowledge regarding spoken content retrieval. The five major emerging directions, as summarized above, are introduced in Sections III, IV, V, VI and VII. Finally, the concluding remarks and the prospects for this area are given in Section VIII.

II. Background Knowledge

In this section we provide background material for spoken content retrieval, primarily for the framework of cascading speech recognition and text information retrieval, but also for useful methods beyond the cascading framework. More complete information can be found elsewhere [1]–[3], [7].

A. Task Description for Spoken Content Retrieval

Spoken content retrieval refers to the task whereby a user enters a query, and the system retrieves the information the user wishes to find from a spoken archive, or a large collection of spoken audio data. The query entered by the user can be either in text or spoken form. The user usually tends to use short key terms as queries [48]. The retrieved items for spoken content retrieval are audio signals (sometimes video). Using spoken queries to retrieve text-based content is another widely studied topic usually referred to as voice search [49], and is out of the scope of this paper.

When the user enters a key term as the query, and the system aims at returning the utterances containing the query term, or the exact occurrence time spans of the query term, the task is referred to as Spoken Term Detection (STD) [50]. Currently, there are major research efforts for spoken content retrieval that focus on this task [51], [52]. Sometimes this task is also referred to as keyword spotting. However, conventionally "keyword spotting" refers to a task with a pre-defined keyword set, or all keywords are known in advance, but for STD the query can be any term, including OOV words. STD can be insufficient though, because a user can prefer to be offered all

spoken content relevant to the query, regardless of whether the query terms are included or not. The task of returning objects semantically related to the query but not necessarily including the query is referred to as semantic retrieval in this article, and which was referred to as spoken document retrieval in some research initiatives [15], [53]–[55]. For semantic retrieval, the retrieval target can be either individual utterances or spoken documents, where the latter includes multiple consecutive utterances with a coherent topic. Topic boundaries of spoken documents in a spoken archive are naturally given in some cases, or can be found by topic segmentation techniques [56].

B. The Framework of Cascading Speech Recognition with Text Retrieval

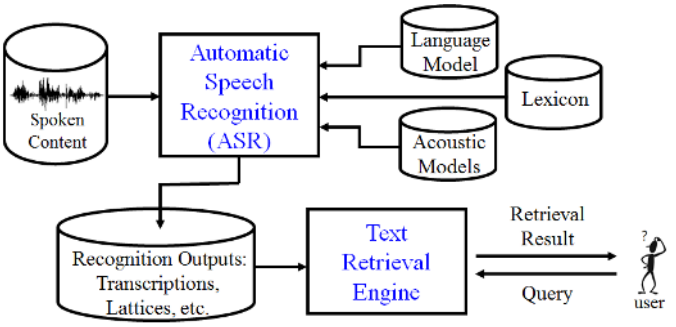

Fig. 1: The basic framework of cascading speech recognition with text retrieval.

An intuitive approach for spoken content retrieval is to use an ASR module to transcribe the spoken content into text first, and then apply text information retrieval on the transcriptions. There are usually two stages cascaded in typical spoken content retrieval systems as shown in Fig 1, for either STD or semantic retrieval. In the first stage (the upper half of Fig 1), the spoken content is processed into transcriptions or lattices by the ASR module, based on an acoustic model, a lexicon, and a language model. In the second stage (the lower half), after the user enters a query, the text retrieval engine searches through the recognition outputs (either transcriptions or lattices) to find the relevant time spans or utterances for STD, or relevant utterances or spoken documents for semantic retrieval. The returned time spans, utterances or spoken documents are assigned scores. Only objects with scores exceeding a threshold are shown to the users, ranked according to the scores.

C. Evaluation Metrics

Because the STD and semantic retrieval scenarios are parallel, most evaluation metrics described here can be used for both tasks, and therefore the *objects* here refer to either time spans or utterances in STD, and utterances or spoken documents in semantic retrieval. The evaluation metrics are separated into two classes [57], as described next.

1) Evaluation of unranked retrieval results: The retrieval performance is evaluated based on the correctness of the retrieved objects only, while the order of the objects in the returned lists is not considered.

Precision, Recall and F-measure are standard metrics. Precision is the fraction of retrieved objects which are relevant, and recall is the fraction of relevant objects which are retrieved. F-measure then integrates precision and recall. Another evaluation metric for unranked retrieval results is the Actual Term Weighted Value (ATWV) [50] whose spirit is very similar to F-measure. ATWV has been widely used to evaluate STD systems today.

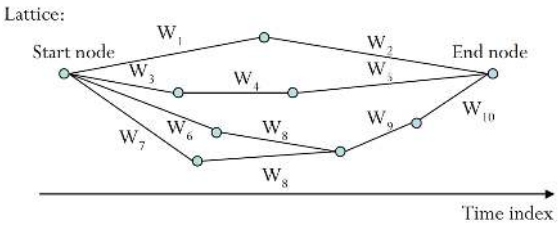

Fig. 2: An example of a lattice.

2) Evaluation of ranked retrieval results: Most commercial search engines display their retrieval results as ranked lists, and the user's satisfaction is highly dependent on the order of the list, so evaluating the order of the ranked list is important.

Precision@N is the precision measure of the top N returned objects. *R-precision* is similar to precision@N, except that N varies for each given query and is set to the total number of relevant objects for the query in the target database. *Mean Average Precision* (MAP) [58] is the **mean** of the *Average Precision* over the testing queries. The average precision for the retrieval results of a query is defined as in (1),

$$Average\ Precision = \frac{\sum_{k=1}^{n} precision(k) rel(k)}{R}, \quad (1)$$

where R is the number of relevant objects for the query in the target database, n is the total number of objects in the returned ranked list, $precision(k)$ is the precision for the top k objects in the list (i.e., Precision@k), and $rel(k)$ is an indicator function which equals to one if the item at rank k is a relevant object, and zero otherwise. The value of MAP can also be understood as the area under the precision-recall curve [57].

D. Lattices

STD based on the one-best transcription is relatively straightforward because the text retrieval engine can search through the transcriptions of the target spoken archive, and the desired objects can be found. However, ASR errors degrade performance, especially if the error rates are high. In order to have better performance, given an utterance, the retrieval engine may consider not only the word sequences with the highest confidence, but all sequences of alternative hypotheses whose confidences are high enough, organized as a lattice as the example in Fig 2, in which each arc W_i represents a word hypothesis. In this way, even though the one-best transcription is not correct, it is possible to find the correct words or word sequences in the lattice. Therefore, lattices are usually preferred for STD, especially when the accuracy in the one-best transcriptions is relatively low [9]. However, too many incorrect word hypotheses in the lattices can also lead to problems. Various approaches, based on posterior probabilities or confidence measures, have been used to try to filter out some of the incorrect word hypotheses, as explained below.

With lattices, an STD system usually returns the time spans of arc sequences a whose hypotheses exactly match the

query Q at the relevant time spans. The confidence scores of the returned time spans are the posterior probabilities of the corresponding arc sequences a [10]–[13], [59], [60]. The posterior probability of an arc sequence a within the lattice of an utterance u, $\mathcal{L}(u)$, is in (2).

$$P(a|u) = \sum_{s \in \mathcal{L}(u), a \in s} P(s|u), \tag{2}$$

where s is an allowed word sequence in $\mathcal{L}(u)$ ($s \in \mathcal{L}(u)$) containing the arc sequence a ($a \in s$), and $P(s|u)$ the posterior of the word sequence s given the lattice $\mathcal{L}(u)$ as in (3).

$$P(s|u) = \frac{P(u|s)P(s)}{\sum_{s' \in \mathcal{L}(u)} P(u|s')P(s')}, \tag{3}$$

where $P(u|s)$ and $P(u|s')$ are the likelihoods for observing the utterance u given the word sequences s and s' based on the acoustic model set, and $P(s)$ and $P(s')$ are the prior probabilities of s or s' given by the language model.

For efficient search, the lattices should be indexed, and the n-gram inverted index is one way to index the lattices [9], [61]–[65]. In the n-gram inverted index, the information about which word or subword n-gram appears in which lattice is stored. Theoretically, the inverted index should contain all possible word or subword n-grams with different lengths, but when n is large, the number of distinct n-grams can be huge. Another approach for indexing the lattice structures is representing the lattices as weighted automata and building an index for all of the possible sub-strings contained in the lattices, which is considered more efficient than the n-gram inverted index [66]. Under this general framework, the index itself is a weighted finite state transducer (WFST) whose inputs are queries represented as text strings, and the outputs are lists of time spans and their scores [66], [67]. When the input query is in audio form, it is also transcribed into a lattice. All the text strings in the query lattice are then used to search over the WFST index, and the final results are the union of the results for each text string. This search process can be efficiently implemented by representing the query lattice as a WFST too, and composing the query WFST with the index WFST [68], [69].

The arcs in the lattices can be gathered into clusters to form sausage-like structures to make the indexing task easier and reduce the memory requirements. Examples of such sausage-like lattice-based structures include Confusion Networks (CN) [70], [71], Position Specific Posterior Lattices (PSPL) [10], [72], [73], and others [74], [75].

E. Out-of-Vocabulary Queries

If a word spoken in the audio is not present in the lexicon of the recognizer, it can never be correctly recognized. Therefore, if a query contains Out-of-Vocabulary (OOV) words, for STD, the retrieval system cannot find the arc sequences of the query even if the retrieval process is conducted on the lattices, since the lattices are constructed with hypotheses of words in the lexicon. Unfortunately, since the users usually enter queries for those they wish to find more information about, the less common words and topic-specific words constitute a good portion of the queries, and many of such words are OOV. Therefore, the percentage of OOV queries was found to be higher than 15% on a real system [76].

Searching over the transcriptions or lattices based on subword units has been a good approach to tackling the OOV problem to some extent. Even though the OOV words cannot be recognized, they can be represented as sequences of subword units, therefore it is possible to find them if the recognition outputs are also represented in subword units [11]–[14], [77]–[86]. These include subword-based lattices in which the arcs are subword hypotheses instead of word hypotheses, or word/subword hybrid lattices, e.g. some arcs in the lattices are word hypotheses, while some others are subword hypotheses [77], [87], [88]. During retrieval, when a query (OOV or not) is entered, it is also converted into a sequence of subword units and then matched with the subword unit sequences in these lattices. Given an OOV query in text form, grapheme-to-phoneme (or letter-to-sound) techniques are needed to estimate the subword sequences for the OOV word [69], [83], [89]–[91], and including multiple alternatives weighted by their confidences is helpful [69], [89]. Subword units may offer better recall for OOV queries as discussed above, very often at the price of lower precision. For example, the subword unit sequence for a query may appear exactly in some utterances consisting of completely different words.

A wide range of subword units has been used in subword-based retrieval, roughly divided into two categories: *linguistically motivated units* (obtained based on some knowledge about the specific language, with good examples include syllables [78], [92], [93], characters (for Mandarin) [78], [92], [93], phonemes [80], or subphone units [84]), and *data driven units* (derived from the corpora utilizing statistical and/or information theoretic principles [14], [79], [81], [82], [94], with the statistical morphs [81], [82], [94] learned from the training corpus as a good example). There are certainly other ways to address the OOV issues in addition to using subword units [79], [95], but left out here for space limitation.

F. Score Normalization

For the unranked evaluation measures such as ATWV (Subsection II-C1), a threshold determining whether an object is considered as relevant is required. However, the characteristics of the queries are usually very diverse. A threshold value good for one query may ruin the performance of another. One way to solve this problem is to estimate a query-specific threshold for each query [13], [96], and another way is to normalize the score distribution of the objects for each query to generate commensurate score distributions for different queries [41], [42], [97]–[99].

G. System Combination

It has been well known in ASR that system combination usually provides improved performance [24]. Because word-based approaches suffer from OOV words and as a result have lower recall, while subword-based approaches result in higher recall but at the price of lower precision, an integration of systems using different units may yield better performance. Therefore, in addition to generating better recognition output with system combination, it is also possible to perform the

combination in the retrieval stage, for example, first generating individual retrieval results from different recognition outputs produced by different ASR systems, and then integrating the retrieval results [78], [100]–[103]. The confidence scores of each object can be the weighted sum of the confidence scores based on each individual recognition output, with weights either learned from training data [101], [102], or optimized based on some evaluation metrics such as MAP [103], [104]. The score normalization introduced in Subsection II-F is also helpful here. It was found that normalizing the confidence scores from different recognition outputs before integrating them may end up with better performance [42].

III. Modified Speech Recognition for Retrieval Purposes

In this section, we present the first major direction: modified speech recognition for retrieval purposes.

A. Motivation

There are several other application tasks in which ASR has been integrated with some downstream processing components in tandem. For example, a spoken language translation (SLT) system is a cascade of ASR and machine translation. In these tasks, although the overall performance heavily depends on the ASR accuracy, the relationship between the overall performance and the accuracy of the ASR module is usually not exactly in parallel. This is reasonable. The traditional word error rate for ASR, which treats all word errors as equally bad, is not necessarily the best measure in calibrating the behavior of the ASR module in these different tasks. Obviously, different word errors have different impact on different tasks (e.g. some function words are important for translation while some others are not); the ASR module minimizing the traditional word error rate therefore inevitably leads to only suboptimal overall performance for different application tasks. This is why it was found that in these cases learning ASR and the downstream subsystems jointly by optimizing the overall performance of the respective application tasks is better than optimizing the ASR module and the downstream processing separately [105]. For example, in SLT, ASR and machine translation have been jointly learned to optimize the bilingual evaluation understudy (BLEU) score [106].

For spoken content retrieval considered here, various studies also pointed out that the traditional word error rate is not always directly related to retrieval performance. First, the terminologies or topic-specific terms constitute a good portion of the queries, so the recognition errors of these terms may have larger influence on retrieval performance, whereas the recognition errors for function words like "the" and "a" have almost no impact. More precisely, the error rates for named-entities were shown to be more correlated with retrieval performance than the normal word error rates treating all recognition errors equally [54], and error rates for those more informative terms weighted by inverse document frequencies were found to be a more accurate indicator for the retrieval performance than the conventional word error rates [107]. Also, it was pointed out that substitution errors have larger influence on retrieval than insertions and deletions because an substitution should be considered as two errors for retrieval [108]; missing the query term in a relevant document may make the document considered as irrelevant, while adding a spurious word into an irrelevant document may make the document considered as relevant. Moreover, ASR errors replacing a word by a semantically dissimilar word were shown to have more impact on retrieval performance than a word with close meaning [109]. Another interesting observation is that although better language models were shown to reduce the ASR error rate, this did not always translate to better STD performance [39], [110], [111]. This is probably because language models tend to bias the decoding towards word sequences frequently appearing in the training data of the language models, but in the training data the terminologies or topic-specific terms often used in the queries are usually rare [111]. In addition, because usually lattices instead of one-best transcripts are used in spoken content retrieval, expected error rate defined over the lattices should be in principle a better predictor of retrieval performance than the error rate of one-best transcriptions [112].

Although it is not easy to try to handle each of the above observations individually, it seems plausible that trying to optimize the ASR module and the retrieval module jointly based on some overall performance for retrieval may provide additional gains as compared to simply minimizing the traditional word error rates for the ASR module alone, as will be discussed more below.

B. Retrieval-Oriented Acoustic Modeling

Three related but different approaches for retrieval-oriented acoustic modeling have been proposed in recent years. They are briefly summarized here.

1) Weighted Discriminative Training: Discriminative training techniques such as minimum classification error (MCE) [113] and minimum phone error (MPE) [30], [114] training have been widely used to obtain better HMM acoustic models, and recently the state-level Minimum Bayes risk (sMBR) [115] training has been shown to be one of the most effective discriminative training methods for acoustic models with deep neural network (DNN). In these methods, a new set of acoustic model parameters θ^* is estimated by maximizing an objective function $F(\theta)$[1],

$$\theta^* = \arg\max_{\theta} F(\theta), \tag{4}$$

$$F(\theta) = \sum_{r=1}^{R} \sum_{s_r \in \mathcal{L}(u_r)} A(w_r, s_r) P_\theta(s_r|u_r), \tag{5}$$

where u_r is the r-th training utterances, w_r the reference transcription of u_r, s_r an allowed word sequence in the lattice $\mathcal{L}(u_r)$, $A(w_r, s_r)$ the accuracy estimated for u_r by comparing s_r with w_r, $P_\theta(s_r|u_r)$ the posterior probability of the path s_r given u_r as defined in (3) (here the acoustic model parameters θ in the ASR module are included as a subscript to emphasize

[1]The MCE, MPE and sMBR can all be formulated as optimizing (5) with different definitions for $A(w_r, s_r)$ [116].

this probability depends on θ[2]), and R is the total number of utterances in the training set. Obviously, maximizing $F(\theta)$ means maximizing the expected accuracy.

$A(w_r, s_r)$ in (5) is usually defined in a way that the accuracies of different words, phonemes or states are equally weighted. However, because optimizing recognition accuracy may not optimize the retrieval performance, the definition of $A(w_r, s_r)$ can be target dependent. In weighted MCE (W-MCE) [117]–[120], the words in a pre-defined keyword set can have higher contributions to $A(w_r, s_r)$ than other words, so the acoustic models can learn to prevent making mistakes when recognizing the words in the keyword set. W-MCE was shown to yield better retrieval performance than the original MCE on Switchboard [118], [119]. With the same principle, when training the DNN acoustic models, by making those states belonging to the pre-defined keywords have more contributions to $A(w_r, s_r)$, the keyword-boosted sMBR [121] is capable of detecting more keywords while reducing false alarms on the NIST Open Keyword Search Evaluation in 2013 (OpenKWS13)[3]. Of course very often in spoken content retrieval the user queries cannot be known beforehand, but there exist ways to find the terms with higher probabilities to be used as queries [122]. Therefore, it is certainly possible to generalize these approaches to other scenarios of spoken content retrieval.

2) Retrieval-Oriented Whole Word Modeling: The above approaches emphasize the keywords, but the optimized acoustic models are also used for other words. A further step forward can be taken by considering the keyword spotting scenario, and training the whole-word models for the keywords only if the keywords have sufficient examples in training data. In this way, the whole-word models can better capture the variability of the keywords and thereby deliver better performance than the conventional phone-level models [123], [124]. A good example in this category is the point process model (PPM) used in keyword spotting, in which the keywords are detected based on the timing of a set of phonetic events (or "landmarks") found in the speech signals [125]–[127].

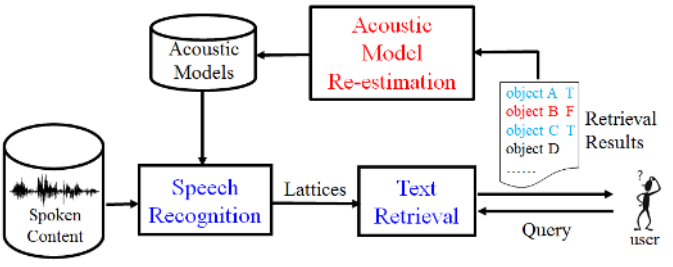

Fig. 3: The framework of re-estimating acoustic model parameters to optimize the overall retrieval performance under relevance feedback scenario.

[2]For $P(s|u)$ in (3), the acoustic model scores $P(u|s)$ is actually computed based on the acoustic model parameters θ. Therefore, $P(s|u)$ in (3) depends on θ. This is not mentioned in Subsection II-D.

[3]An overview of NIST OpenKWS13 can be found at: http://www.signalprocessingsociety.org/technical-committees/list/sl-tc/spl-nl/2013-08/sltc-newsletter-august-2013-overview-of-the-nist-open-keyword-search-2013-evaluation-workshop/

3) Retrieval-oriented Acoustic Modeling under Relevance Feedback Scenario: Relevance feedback [128] well used in text retrieval is useful to integrate the ASR and retrieval modules as a whole and optimize the overall retrieval performance, rather than considering them as two cascaded independent components [129]–[132], as shown in Fig. 3. When a query is entered by the user, the system offers a ranked list of retrieved objects to the user. If the user gives some feedback to the system, for example, selecting items 1 and 3 as shown in Fig. 3 (implying relevant) but not item 2 (probably implying irrelevant), a new set of acoustic models can then be re-estimated on-line based on the feedback. Because the scores used for ranking the objects depend on the acoustic models, the objects below item 3 not yet viewed by the user can thus be re-ranked. In this way, the acoustic models can be "adapted locally" considering the specific query and the corresponding feedback entered by the individual user, resulting in "query-specific" acoustic models, to be used for the unlimited number of acoustic conditions for the spoken content. This framework has been successfully applied on STD with utterances as the retrieval target [129]–[132] as briefly explained below.

In STD with utterances as the retrieval target, when the query Q is entered, all utterances u in the spoken archive are ranked according to a confidence score $S_\theta(Q, u)$, where θ is the set of acoustic model parameters. The expected frequency of the query Q in the utterance u, $E_\theta(Q, u)$, is an example of $S_\theta(Q, u)$.

$$S_\theta(Q, u) = E_\theta(Q, u) = \sum_{s \in \mathcal{L}(u)} N(Q, s) P_\theta(s|u). \quad (6)$$

Equation (6) is parallel to the core part of (5), except that $A(w_r, s_r)$ in (5) is replaced by $N(Q, s)$, the occurrence count of the query Q in an word sequence s which the lattice $\mathcal{L}(u)$ allows. Given positive and negative (or relevant and irrelevant) examples for a certain query Q from the user relevance feedback as explained above, the system estimates a new set of acoustic model parameters θ^* by maximizing an objective function very similar to (4) and (5) as in Subsection III-B1, but with different definitions of $F(\theta)$ as explained below. With the new set of acoustic models θ^*, (6) is modified accordingly[4], and the retrieved results not yet viewed by the user are re-ranked.

The objective function in (4) can be the sum of the differences between all positive and negative example pairs here,

$$F_1(\theta) = \sum_{u_t, u_f} [S_\theta(Q, u_t) - S_\theta(Q, u_f)], \quad (7)$$

where u_t and u_f are respectively positive and negative example utterances. By maximizing (7) as in (4), the system tries to learn a new set of models θ^* which better separates the scores of relevant and irrelevant utterances.

Also, it has been shown that maximizing $F_2(\theta)$ below is equivalent to maximizing a lower bound of the retrieval

[4]With the new acoustic models θ^* to update $S_\theta(Q, u)$ in (6), only $P_\theta(s|u)$ in (6) have to be changed without generating new lattices, so updating $S_\theta(Q, u)$ on-line is not computation-intensive [129].

performance measure MAP in Subsection II-C2 [133], [134]:

$$F_2(\theta) = \sum_{u_t, u_f} \delta(u_t, u_f), \tag{8}$$

where $\delta(u_t, u_f)$ is 1 if $S_\theta(Q, u_t) > S_\theta(Q, u_f)$, but 0 otherwise. $F_2(\theta)$ hence represents the number of positive/negative example pairs for which the score of the positive example is greater than that of the negative example. In addition, it will be beneficial to include the large number of unlabelled data in the training process by assuming the unlabelled objects are irrelevant. $F_3(\theta)$ below realizes the above idea.

$$F_3(\theta) = F_2(\theta) + \rho \sum_{u_t, u_{no}} \delta(u_t, u_{no}), \tag{9}$$

where u_{no} is an unlabelled utterance within the returned list, and ρ is a weighting parameter. Experimental results showed that all the object functions in (7) to (9) improved the retrieval performance, $F_2(\theta)$ was superior to $F_1(\theta)$, while $F_3(\theta)$ further outperformed $F_2(\theta)$ [129].

C. Retrieval-Oriented Language Modeling

In keyword spotting, it was found that boosting the probabilities of n-grams including query terms by repeating the sentences including the query terms in the language model training corpora improved the detection performance in the evaluations of DARPA's Robust Automatic Transcription of Speech (RATS) program [135] and the NIST OpenKWS13 Evaluation [136]. Similar concept was also used in neural network based language models (NNLM) [28], whose input is a history word sequence represented by a feature vector, and the output is the probability distribution over the words. NNLMs are trained by minimizing an objective function representing the differences between words in the training corpus and the output distributions given their history word sequences. It was found that NNLM decreased the word error rate and perplexity, but may hurt STD performance at the same time [39], so new training strategy for NNLM was proposed [47]. In the new strategy, if a word is less frequent in the training corpus (which has higher probability to be the query term), in the objective function the difference measure obtained between this word and the output distribution of NNLM was weighted, and thus the NNLM learned to prevent making mistakes on the words with larger weights. It was found that this training strategy improved the STD performance on Vietnamese in the NIST OpenKWS13 Evaluation [47]. NNLMs trained in this way were also found to offer higher word error rates and perplexity compared with the conventional NNLM, which is another evidence to support that ASR module specially designed for spoken content retrieval is a reasonable direction.

D. Retrieval-Oriented Decoding

It has been proposed that the search with OOV queries can be achieved in two steps [110]. In this framework, each utterance has a word-based and a subword-based lattices. When an OOV query is entered, in the first step, a set of utterances which possibly contain the OOV query is obtained by searching over the subword-based lattices. Decoding these utterances with a new lexicon including the OOV terms in the query and then searching over the new lattices thus obtained can yield more precise results compared to the subword-based retrieval, but generating new lattices on-line is not tractable. Therefore, instead of generating new lattices, this approach inserts the word arcs whose hypotheses are the OOV terms into the word-based lattices. The time spans of these arcs are those obtained in the first step. Then the word-based lattices are re-scored to obtain the acoustic likelihoods and language model scores of the new arcs, and the second step retrieval is conducted on the re-scored lattices. Here the system only re-scores the existing lattices instead of decoding the utterances, so this framework can be realistic. For OOV queries, this framework achieved 8.7% relative improvement over subword-based retrieval on MIT iCampus lecture set [110].

Sometimes even in-vocabulary query terms in the utterances cannot be found in the lattices, because the hypotheses for those in-vocabulary query terms have relatively low language and/or acoustic model scores, and therefore they are pruned when generating the lattices. This is especially serious for keywords which is rarely used and thus have low language model scores. Subword-based retrieval may address this issue as mentioned, but the retrieval results based on subwords can be noisy with poor precision. Another solution to this problem is to increase the depth of the word lattices, but this may seriously increase the computation and memory requirements. A more realistic solution is to give different words different pruning thresholds during decoding [97], [135][5]. By giving the interested keywords much lower pruning thresholds compared with normal terms, this method obtained better performance than the subword-based solution [97].

E. Retrieval-Oriented Confusion Models

Some effort has been made to model the occurrence of the recognition errors in a systematic way, referred to as confusion models here, and to try to optimize such models to have better retrieval performance. There can be at least three ways to achieve this goal: *Query transformation* [95], [137], [138] (to transform the word or subword sequence of each query into the sequences that the query tends to be mis-recognized to, and the new set of sequences are used to retrieve the lattices), *Spoken Content transformation* [139]–[141] (to transform the recognition output for the spoken content instead of the query), and *Fuzzy match* [142]–[146] (defining a distance between different word or subword sequences, and the lattices containing word or subword sequences sufficiently close to the query being retrieved).

In all the above, a confusion model describing how the confusion of a word or subword sequence is to the other is needed. Usually this model is represented as a P by P matrix, where P is the number of subword units considered[6]. In this matrix, the value of the element at i-th row and j-th column indicates the probability that the i-th subword unit may be misrecognized as the j-th subword unit (therefore this matrix is not symmetric). The confusion between the word or subword sequences can then be obtained. It has been proposed to learn such confusion models or matrices by optimizing the retrieval

[5] Also called white listing [97] or keyword-aware pruning [135].

[6] Although the confusion of subword n-grams or words can be considered, they are not widely used because of lack of training data.

evaluation metrics using a set of training queries and the corresponding audio [137], [139], [140]. For the experiments of STD on Fisher corpus, the model thus learned yielded 11% relative improvements in terms of Figure of Merit (FOM) over the baseline without transformation [140].

F. Jointly Optimized Discriminative Model Integrating Recognition and Retrieval

A very interesting different approach is to try to define a function $S_J(Q,u)$ which can map the acoustic features of an utterance u and a query Q to a confidence score $S_J(Q,u)$, to be used to rank the utterances just as $S_\theta(Q,u)$ in (3). In this way the speech recognition and retrieval are integrated in a single function $S_J(Q,u)$, which can be optimized by learning from some overall retrieval goal. Encouraging results have been obtained on STD with some preliminary approaches along this direction [147]–[151]. In these approaches, the above confidence score is formulated as in (10).

$$S_J(Q,u) = arg\ max_{h\in u} w\cdot\phi(Q,h), \tag{10}$$

where h is any signal segment in the utterance u ($h \in u$), $\phi(Q,h)$ is the vector of a set of features describing the likelihood that Q appears in h (explained below), w is a weight vector to be learned from training data, and $w\cdot\phi(Q,h)$ is interpreted as the confidence that Q appears in h. In (10), the score of the most confident signal segment h, that is, the signal segment h with the largest $w\cdot\phi(Q,h)$ among all possible h, is the confidence score for the utterance u. The feature vector $\phi(Q,h)$ can include various kinds of information useful for STD, such as the outputs of the phoneme classifiers based on different models (e.g. gaussian mixture model, recurrent neural networks, etc.) and the outputs of articulatory feature classifiers [147]–[151]. Although there is an exponential number of possible segments h in u which may make (10) intractable, with carefully designed feature vector $\phi(Q,h)$, dynamic programming algorithm can efficiently solve (10) [147]. With a set of training queries and their relevant and irrelevant utterances, w can be learned to maximize the evaluation metrics of STD. Because only a single vector w is used here to model both ASR and retrieval in a very different framework, it may not be easy to compare directly these approaches with conventional approaches using state-of-the-art ASR modules. However, these approaches have been shown to work very well in the setting of very limited training data, for which it may not be feasible to train an ASR module reasonably well [147], [148]. For the experiments on Switchbroad, the approach based on (10) outperformed the baseline with an ASR module using phone-based HMMs when the training audio size is over a range from 500 to 5000 utterances [148].

IV. Exploiting the Information Not Present in Standard ASR Outputs

In this section, we present the second major direction: exploiting information not present in ASR outputs.

A. Motivation

In addition to the posterior probabilities from the lattices in (2) to be used as the confidence scores for retrieval, other useful cues for confidence score estimation were found in the lattices. One example is the context of the retrieved objects within the lattices [152]–[156]. Another example is the outputs of an OOV detector which detects the presence of an OOV word by analyzing the score distributions of the arcs in the lattices [157]. If the input query is OOV, and a time span is detected as an OOV word, the corresponding confidence score can be boosted [158].

On the other hand, when speech signals are decoded into transcriptions or lattices in ASR, much of useful information are no longer present, for example, the temporal variation structure of the signals. Therefore, when the retrieval processes were applied on top of the ASR outputs, it is a good idea to consider if the information not present in ASR outputs can be used in enhancing the retrieval performance. A good example is to include prosodic cues, and another series of work tried to perform query-specific rescoring using such information.

B. Incorporating Prosodic Cues

Duration related cues have been shown useful, such as the duration of the signal segments hypothesized to be the query divided by the number of syllables or phonemes in the query (or the speaking rate), and the average duration of the same syllables or phonemes in the target spoken archive [149], [150], [159]–[162]. This is because extremely high or low speaking rate or abnormal phoneme and syllable durations may imply that the hypothesized signal segment is a false alarm. The maximum, minimum and mean of pitch and energy in hypothesized signal segments were also found to be useful [159], [160], since extreme values of pitch and energy usually cause more ASR errors [163], thus helpful to identify the false alarms. Moreover, it was found that the results of landmark and attribute detection (with prosodic cues included) can reduce the false alarm [164]. Thorough analysis for the usefulness of different kinds of cues also indicated that the cues related to duration are very useful cues [159], [160].

To integrate the different cues such as those mentioned above, regardless of whether obtained in ASR outputs or not, the STD problem has been formulated as a binary classification problem [159], [160], [165]. Each candidate object x is represented as a feature vector $f(x)$, with each component in $f(x)$ for a cue (e.g. posterior probabilities, confidence scores, duration or pitch related features, etc.). Then a classifier can learn to classify those objects x to be true or not based on its feature $f(x)$ if a set of training queries and their associated true/false examples are available. Such classifiers can be any kind of binary classifiers including support vector machines (SVMs), deep neutral networks (DNNs) and so on.

C. Query-specific Rescoring based on Pseudo-relevance Feedback (PRF)

In STD after a query is entered, the system can focus on locating the time spans of only the specific query terms in the spoken archive, not any other phoneme sequences or any other terms or words. This implies the possibility of learning query-specific rescoring approaches; i.e., the goal is focused on simply exploiting the specific acoustic characteristics of a given query. This is quite different from the conventional

ASR, for which the occurrence of all possible phonemes and words have to be considered. The concept of such query-specific rescoring makes it possible to consider the acoustic characteristics of the specific query which is not present in ASR transcriptions. This is easier to achieve (the scope is limited) than to consider the situations for all possible phonemes or words. Although this sounds impractical since we need training data for each query, but can actually be realized with pseudo-relevance feedback (PRF). Also, different from ASR tasks in which only the input utterance is focused on, for the STD tasks, all signal segments hypothesized to be the query in the whole target spoken archive can be explored.

Pseudo-relevance feedback (PRF), also known as blind relevance feedback, has been successfully applied on different retrieval domains including those for text [166]–[170], image [171], [172] and video [173]–[175]. When applied in STD scenario, it can be used to obtain a query-specific training set to train query-specific rescoring approaches. The framework is shown in Fig 4. A first-pass retrieval is performed first using some scores such as those in (2) or (6), with results not shown to the user. A small number of retrieved objects with the highest scores is then taken as "pseudo-relevant", and sometimes some objects with the lowest scores as "pseudo-irrelevant" in addition. Not all these examples are labelled correctly, but they should have signal characteristics reasonably similar or dissimilar to the possible acoustic characteristics of the query since they are found from the whole target spoken archive in the first-pass retrieval. These positive and negative examples are then used to train a query-specific rescoring model to rescore and re-rank the objects in the first-pass retrieved list. The system finally displays the re-ranked results to the user. Several ways to realize this query-specific rescoring are in the next subsection.

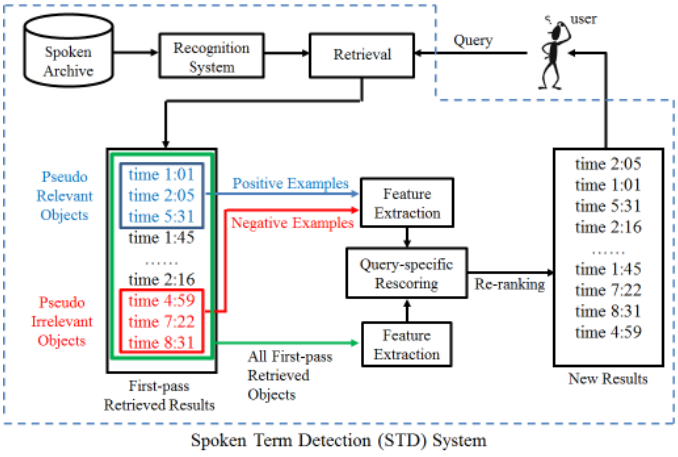

Fig. 4: The pseudo-relevance feedback (PRF) framework of training query-specific rescoring approaches for spoken term detection (STD).

D. Different Approaches for the Query-specific Rescoring

1) Query-specific Detector: One way to realize query-specific rescoring is to learn the query-specific detectors[7], whose inputs are the hypothesized regions for a specific query, and the outputs are whether the hypothesized regions are correct [177], [178]. Here the hypothesized region is defined as the segment of acoustic vectors (e.g. MFCCs) corresponding to an arc sequence a in the lattice, which has the highest confidence score among all arc sequences in the lattice with their hypotheses being the query Q as shown in the left half of Fig. 5. In the right half of Fig. 5, a hypothesized region is divided into a sequence of divisions based on the HMM state boundaries obtained during the lattice construction. Each division is then represented by a vector which is the average of the acoustic vectors in it. All these averaged vectors in a hypothesized region are then concatenated to form a feature $f(x)$ for the hypothesized region x. For l-state phoneme HMMs and a query term Q including m phonemes, the dimensionality of such a feature vector $f(x)$ is $m \times l$ times the dimensionality of the acoustic vectors. The feature vector $f(x)$ thus capsules the acoustic characteristics of the hypothesized region. Note that much of such information is lost when the acoustic vector sequence is transformed into the lattice by ASR. These features for the positive and negative examples can then be used to train an SVM or DNN classifier. It was shown that the re-ranked results yielded significant improvements over the first-pass results on both course lectures and broadcast news with SVM [177], and on TIMIT corpus with DNN [178]. This approach can be equally applied when the retrieval objects are utterances. The only difference is that the first-pass retrieved results in Fig. 4 are the lattices in the left half of Fig. 5 for the latter, but directly the hypothesized regions for the former. Below we always assume the retrieval target is the hypothesized region for simplicity, although all mentioned are equally applicable for utterances.

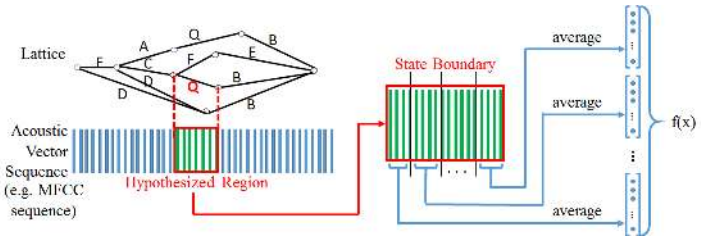

Fig. 5: Feature vector representations. Left half: a hypothesized region of the query term Q from the lattice. Right half: the feature vector $f(x)$.

2) Exemplar-based Approach: Exemplar-based approaches have been identified as a new paradigm which may enhance the conventional HMM-based ASR [31]. The limited number of parameters in HMMs are inadequate for representing the fine details of the training audio signal set. Greatly increasing the number of parameters may make the model over-fitted with the training data. It was thus proposed to use the similarity between the utterances being considered and a set of word or phoneme examples in transcribing the utterances. Such approaches were shown to be able to improve the conventional HMM-based ASR, and referred to as the exemplar-based or template-based ASR [31]. Such information as temporal structures or trajectories of signals can be exploited in this way [31], hopefully having the potential to address the deficiency of conventional HMMs.

[7]The concept is similar to "utterance verification" or "confidence score estimation" [176], although the scenarios may not be the same.

In STD, the above exemplar-based techniques have also been considered. For an input query, assume some signal segments corresponding to the query terms and some others corresponding to other terms but easily mis-recognized as the query terms by conventional ASR techniques are available as positive and negative examples. These examples can help to rescore and re-rank the hypothesized regions obtained in the first-pass retrieval. For example, those hypothesized regions more similar to the positive examples than the negative examples are more likely to be truly relevant. This is formally formulated as below. Given N training examples for a given query Q, $\{x_i^Q\}_{i=1}^N$, each has a label $y_i \in \{-1, 1\}$, where 1 for positive and -1 for negative examples. The confidence for a hypothesized region x being the query Q can then be represented as (11).

$$S(Q, x) = \sum_{i=1}^{N} w_i W(x, x_i^Q), \tag{11}$$

where $W(x, x_i^Q)$ represents the similarity between the hypothesized region x and the example x_i^Q, and w_i are the weights for example x_i^Q. Intuitively, w_i should be close to the label y_i, i.e., 1 for positive and -1 for negative examples. Practically the weights $\{w_i\}_{i=1}^N$ can be learned [179][8]. There are also various ways to obtain the similarity $W(x, x_i^Q)$ between a hypothesized region and an example, both represented as acoustic vector sequences. One way is to represent the acoustic vector sequences by fixed length feature vectors as in Fig 5 [180], and then compute the similarity between the two fixed length feature vectors. Another way is to use dynamic time warping (DTW) to evaluate the similarity between two acoustic vector sequences with different lengths [181]–[183]. DTW will be further discussed later on in Section V. This approach was shown to yield significant improvements on both course lectures [181] and broadcast news [184][9]. This approach was also shown to offer improvement additive to the retrieval-oriented acoustic modeling under relevance feedback scenario in Subsection III-B3 [129].

E. Graph-based Approach

The query-specific rescoring based on PRF in the last subsections can be taken one step further. It is reasonable to expect that globally considering the similarity structure among all hypothesized regions obtained in the first pass, rather than relying on the assumptions of examples in PRF, can better re-rank the hypothesized regions. This can be formulated as a problem using graph theory [180]–[183], [185]. As shown in Fig. 6, for each query Q a graph is constructed, in which each node x represents a hypothesized region for the query Q from the first pass, and two nodes are connected if the similarity between the two corresponding hypothesized regions is high. The edge weights $W(x, x')$ between the nodes x and x' are the similarity between them, as in the left half of Fig. 6, which can be estimated with different approaches including DTW.

One way to exploit the graph structure in Fig. 6 is to use the minimum normalized graph cut [180]. The minimum normalized graph cut [186] splits the nodes in a graph into two disjoint groups, where the inter-group edge weights are low, and the inner-group edge weights are high. Since the true hypothesized regions corresponding to the query term should have relatively similar acoustic characteristics, or be strongly connected on the graph, so minimum normalized graph cut can separate true and false hypothesized regions into two groups. To determine which group is for the true hypotheses, the system samples one node and asks the user to label whether it is true. Minimum normalized graph cut also provides each node a score representing the tendency of belonging to the two groups [180], so the hypothesized regions can be ranked according to this score.

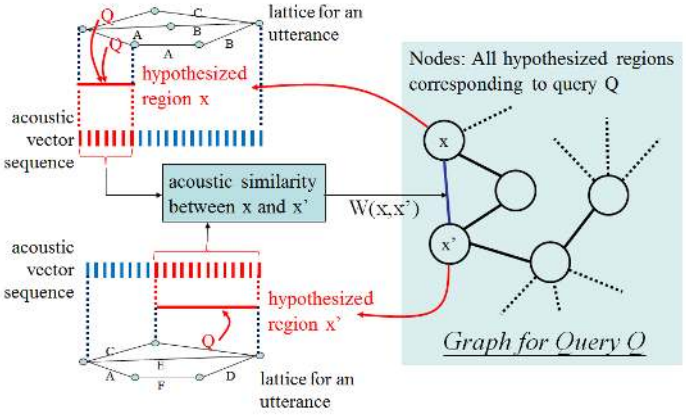

Fig. 6: The graph constructed for all hypothesized regions obtained in the first-pass retrieval with a query Q. Each node in the graph represents a hypothesized region, and the edge weights represent the acoustic similarities between the nodes.

Another way to exploit the graph structure is using the *random walk* [181]–[183], [185], which does not use any labelled data. The basic idea is that the hypothesized regions (nodes) strongly connected to many other hypothesized regions (nodes) with higher/lower confidence scores on the graph should have higher/lower scores. The original confidence scores of the hypothesized regions, which is based on the posterior probabilities from the lattices, therefore propagate over the graph, and then a set of new scores for each node are obtained accordingly. This approach is similar to the very successful PageRank [187], [188] used to rank web pages and compute an importance score for each page. Similar approaches have also been found useful in video search [189], [190] and extractive summarization [191], [192], in which the similarities are used to formulate the ranking problem with graphs.

In this approach, given the graph for a query Q, each hypothesized region x is assigned a new confidence score evaluated with graph $S^G(Q, x)$,

$$S^G(Q, x) = (1 - \lambda) S(Q, x) + \lambda \sum_{x' \in N(x)} S^G(Q, x') W'(x', x), \tag{12}$$

where $S(Q, x)$ is the original confidence score from lattices such as those in (2) or (6), $N(x)$ is the set of nodes having connection with x, x' is a node in $N(x)$, and $W'(x', x)$ is the edge weight between x' and x, but normalized over all edges

[8]This approach is referred to as kernel-based binary classifier [179].

[9]In these experiments, the weights w_i were simply set as $w_i = 1, -1$ for positive and negative examples.

connected to x':

$$W'(x',x) = \frac{W(x',x)}{\sum_{x'' \in N(x')} W(x',x'')}, \quad (13)$$

where $W(x',x)$ is the similarity between x' and x. λ in (12) is an interpolation weight. Here (12) implies $S^G(Q,x)$ depends on two factors, the original scores $S(Q,x)$ in the first term and the scores propagated from similar hypothesized regions x' in the second term. The weight normalization in (13) implies the score of node x' is distributed to all nodes x'' having connection with x'. Although it is possible to use $S^G(Q,x)$ in (12) for ranking directly, integrating $S^G(Q,x)$ with the original score $S(Q,x)$ again by multiplying them was shown to offer even better performance.

The graph-based approach with random walk was shown to yield dramatic improvements on retrieval over a spoken archive produced by a single speaker, for example, course lectures. This is reasonable because for the same speaker the similarity among realizations of the same query terms are relatively high, based on which the random walk is able to very well enhance the confidence scores. In the experiments on lectures for a course taught by a single instructor, 21.2% relative improvement for speaker independent recognition was obtained.It also yielded 13% relative improvement for a set of OOV queries on audio recordings of McGill course lectures [193] with several speakers [185], and 6.1% relative improvements on broadcast news with many speakers [184]. The graph-based approach with random walk was also shown to outperform the exemplar-based approach with examples from PRF [181]. This is because the exemplar-based approach only considers those information for objects most confident to be relevant or irrelevant, whereas the graph-based approach globally considers all the objects retrieved in the first pass.

V. Direct Matching on Acoustic Level without ASR

In this section, we present the next major direction: direct matching on acoustic level without ASR.

A. Motivation

There can be either text or spoken queries. Entering the queries in spoken form is attractive because this is the most natural user interface. Smartphones and hand-held or wearable devices make spoken queries an even more natural choice. Retrieving spoken content with spoken queries is also referred to as *query-by-example*. In principle, *query-by-example* is more difficult than using text queries because both the content and the queries are to be recognized and include recognition errors. Since the spoken queries are usually short without context information, often including OOV words and entered under uncontrolled conditions, resulting in relatively low recognition accuracies. However, the spoken queries also offered a new direction which was never possible for text queries; that is, because both the content and the queries are in speech, it becomes possible to match the signals directly on acoustic level without transcribing them into phonemes or words. Spoken content retrieval becomes possible without ASR.

Giving up ASR inevitably gives up much useful information offered by ASR, but also implies all the difficult problems ever considered for retrieval with ASR are automatically bypassed or eliminated. Such problems include the difficult problems of OOV words, recognition errors, low accuracies due to varying acoustic and noisy conditions, as well as the need for reasonably matched corpora (and annotating them) for training the acoustic/language models to transcribe the spoken content. For low-resourced languages with scarce annotated data, or languages without written forms, recognition seems even far from possible. In particular, it makes great sense to bypass the need for the huge quantities of annotated audio data for supervised training of acoustic models. This is why this direction is also referred to as unsupervised retrieval of spoken content, or unsupervised STD. This direction exactly matches the target of the Spoken Web Search (SWS) task [194]–[197][10], a part of the MediaEval campaigns [198], and some results in the program will be mentioned here. A complete overview of the approaches developed in SWS in 2011 and 2012 is available [199].

The work along this direction can be roughly divided into two categories: DTW-based and model-based. The former compares the signals by template matching based on the very successful approach of dynamic time warping (DTW), while the latter tries to build some models for the signals and wish to benefit from the nice properties of acoustic models.

B. DTW-based Approaches

The most intuitive way to search over the spoken content for a spoken query is to find those audio snippets that sound like the spoken query by directly matching the audio signals. Since the audio events in speech signals can be produced at different speeds with different durations, the spoken content and the spoken query are hardly aligned at the same pace. The dynamic time warping (DTW) approach [200] was invented to deal with exactly such problems. DTW allows a nonlinear mapping between two audio signals or feature vector sequences, namely the query sequence and the document sequence, and produce a minimum distance between the two based on an optimal warping path found by dynamic programming.

Assume we are given a query sequence $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_{|\mathbf{X}|})$ and a document sequence $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_{|\mathbf{Y}|})$, where $\mathbf{x}_i$ and $\mathbf{y}_i$ are frame-based acoustic feature vectors (e.g. MFCCs). Let $\rho(\mathbf{x}_i, \mathbf{y}_j)$ be the pairwise distance between the acoustic feature vectors $\mathbf{x}_i$ and $\mathbf{y}_j$, also referred to as local distance. The goal of DTW is to find a warping path on the (i,j)-plane as in Fig. 7 with the lowest total distance accumulating all $\rho(\mathbf{x}_i, \mathbf{y}_j)$ along the path from $(1,s)$ to $(|\mathbf{X}|, e)$; this represents the matching of $\mathbf{X}$ to $(\mathbf{y}_s, \ldots, \mathbf{y}_e)$. For the circled path in Fig. 7 $s = 1$ and $e = 10$. The spoken documents can than be ranked based on this lowest distance.

1) Segmental DTW: The classical DTW algorithm simply tries to match two sequences $\mathbf{X}$ and $\mathbf{Y}$ primarily end-to-end [201], different from the task considered here. Because the spoken query $\mathbf{X}$ is usually only a small part in a spoken

[10]It was renamed as "Query by Example Search on Speech Task" (QUESST) in 2014.

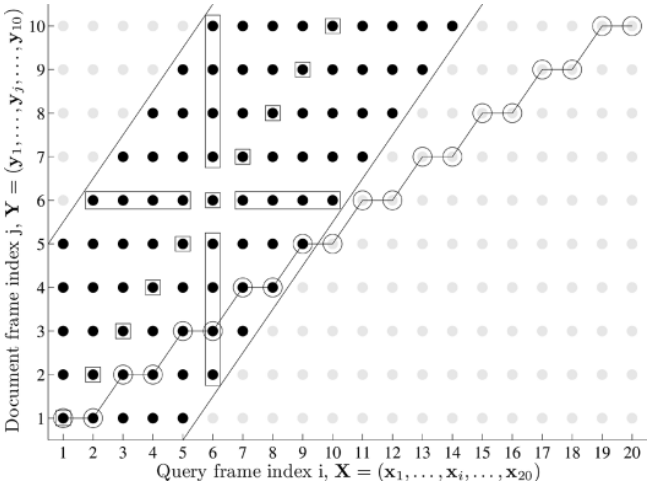

Fig. 7: The matching of query $\mathbf{X}$ to document $\mathbf{Y}$ ($|\mathbf{X}| = 20$, $|\mathbf{Y}| = 10$). With segmental DTW, the diagonal band starting from $(1, 1)$ with bandwidth set to 4 gives each point on the diagonal path $i = j$ (squares) an allowance of 4 points on both sides for both i and j (rectangles), and therefore confines the warping paths to the darkened region. Slope-constrained DTW permits a warping path (circles) that goes from $(1, 1)$ to $(20, 10)$ if each frame in query $\mathbf{X}$ is allowed to match at most 2 frames in document $\mathbf{Y}$, and vice versa, but there is no such path in segmental DTW with bandwidth equal to 4.

document $\mathbf{Y}$, we need to locate the spoken queries in the documents. This is why segmental DTW is needed.

The segmental DTW was first used in unsupervised discovery of speech patterns from spoken documents [202], [203], but it can also be used here. The naming of "segmental" refers to partitioning the (i, j)-plane into several overlapping diagonal bands each with a different starting point and a bandwidth. For example, in Fig. 7, a diagonal band starting from $(1, 1)$ with bandwidth 4 is shown in dark points. Segmental DTW then iterates through all diagonal bands, finding one optimal warping path with minimum accumulated distance within each diagonal band. Each diagonal band serves as a candidate location of the spoken query, with allowed temporal distortion defined by the width of the band.

2) Subsequence DTW and Slope-constraints: In segmental DTW, the spoken query and the matched signal segment in the spoken document can have lengths differ at most by the width of the diagonal band. It works fine with signals of similar speaking rates, but could be problematic in real world applications. Observations over the audio data indicate that the length of the spoken query can be more than twice as long as the same term in the spoken content such as broadcast news, specially because users tend to slow down their voice query to make the pronunciations clear [204]. When the speaking rates differ widely, the longer the query, the wider the duration difference. In order to handle this problem, subsequence DTW [201], [205]–[207] gives up the diagonal bands of segmental DTW, but considers the best match between the query sequence and every possible subsequence of the document exhaustively by dynamic programming. This approach turned out to be very useful.

Another approach is to apply to the local distance a penalty multiplicand, which exponentially grows with the number of query frames matched to the same document frame [208], or the local slope of the warping path. Similar penalty is applied when multiple document frames were mapped to the same query frame, but the collected distance for the same query frame is further normalized by the count of the corresponding document frames; this ensures the final accumulated distance is equally contributed by every frame in the query regardless of how many frames was mapped to each of them. A similar approach, slope-constrained DTW, was also proposed [204]. In this approach, each frame in query $\mathbf{X}$ is allowed to match at most a certain number of frames in document $\mathbf{Y}$, and vice versa. For example, as shown in Fig. 7, the warping path (circles) is for slope-constrained DTW that each frame in document $\mathbf{Y}$ is allowed to match at most 2 frames in query $\mathbf{X}$. It was shown that such slope-constrained DTW offered similar retrieval performance to segmental DTW, but greatly outperformed segmental DTW when the speaking rate difference is large [204].

3) Acoustic Feature Vectors and Distance Measures used in DTW: It is important how one specifies $\mathbf{x}_i$, $\mathbf{y}_j$ and evaluates the local distance $\rho(\mathbf{x}_i, \mathbf{y}_j)$. The simplest way is to use MFCCs for $\mathbf{x}_i$, $\mathbf{y}_j$ and Euclidean distance for $\rho(\mathbf{x}_i, \mathbf{y}_j)$, but this approach implies that MFCC sequences with large distances are from different terms, which is not necessarily true. The posteriorgrams (vectors of posterior probabilities for a set of classes) have been used by most work to incorporate the acoustic feature distribution into distance measures.

Gaussian posteriorgrams have been used for $\mathbf{x}_i$ and $\mathbf{y}_j$ [209]–[211]. To generate Gaussian posteriorgrams, a Gaussian mixture model is trained, and each signal frame is then represented by the vector of the posterior probabilities of being generated from each Gaussian. The Gaussians can be viewed as anchor points in the MFCC space, and the posterior probability translates to the normalized distance to the mean of each Gaussian. It was also proposed to use an multilayer perceptron (MLP) to transform the MFCCs into phonetic posteriors [208]. Though supervised MLP training was needed in this way, the MLP trained from another annotated corpus (probably in a different language) can be used instead [206], [212], [213] because the MLP's phone posterior output can always serve as features, even for a phone set different from that for the target audio. The bottle-neck features derived from MLP can further be used to generate Gaussian posteriorgrams [214], [215]. The local distance for such posteriorgrams, Gaussian or phonetic, is very often defined as the negative log of the inner product;

$$\rho(\mathbf{x}_i, \mathbf{y}_j) = -\log(\mathbf{x}_i \cdot \mathbf{y}_j). \tag{14}$$

Other concepts of defining the features were also proposed, including model posteriorgrams (will be mentioned again in Subsection V-E) [216], [217], RBM posteriorgrams [218] and intrinsic spectral analysis (ISA) features [219], [220]. The performance comparison for spoken term discovery task was reported for different feature representations and distance measures [221].

C. Speed-up Approaches for DTW

One of the major issues of DTW is the high computation demand. One way to speed up DTW is to parallelize the task by distributing the workload to multiple processors on multi-core servers [222] or graphics processing units (GPUs) [223]. The other way is to develop some speed-up approaches to reduce the computation requirement of DTW, sometimes at the price of degraded performance. In most cases, the performance can be preserved by performing a second pass rescoring using DTW on the reduced search space after the first pass filtering using speed-up approaches. The speed-up approaches are discussed below.

1) Segment-based DTW: Both the spoken query and the spoken documents are divided into segments of acoustically similar frames $\vec{\mathbf{x}}_i$ and $\vec{\mathbf{y}}_j$, where $\vec{\mathbf{x}}_i$ is the i-th segment of the query, and $\vec{\mathbf{y}}_j$ is the j-th segment of the document, each consisting of a number of frames. Hence, the DTW is reduced to finding a warping path in the (i, j)-plane of segments based on a carefully designed local distance of $\rho(\vec{\mathbf{x}}_i, \vec{\mathbf{y}}_j)$. Everything for the segment-based DTW is very similar to the original frame-based DTW, except the path searching time is reduced significantly [224]. The signal segments can be generated using the hierarchical agglomerative clustering (HAC) approach [225] by minimizing the total variance greedily when merging two adjacent clusters into one in each iteration. This approach provides a much faster, though coarser, first-pass filtering for selecting possible hypothesized utterances to be used in second-pass rescoring using frame-based DTW.

2) Lower bound estimation: This approach has been proposed for DTW-KNN (K-nearest neighbor) search [226], and used in segmental DTW for STD [227], [228]. The basic idea is to compute the lower bound of the local distance for each frame in the query off-line, which can be achieved by taking the maximum value of the posteriorgram in the window without knowing the query frame. Since the goal is to find the K-nearest snippets in the spoken archive, the snippets are sorted by their lower bound estimation. Starting from the one with the least lower bound, snippets are rescored again and put into a priority queue of size K. The rescoring process hits an early break when the next snippet to run the DTW has higher lower bound than the K-th smallest DTW distance in the queue.

3) Indexing the Frames in the Target Archive: In addition to the path search, another heavy computational cost is from the local distance calculation. To find the path on the (i, j)-plane, the local distance for almost every pair of a frame in the query and that in the spoken documents in the target archive is to be computed. This requires a great deal of computation, even though some frames in the archive are very dissimilar to others. A more clever way is to try to index all the document frames in the target archive. Then for each frame in the query, only those frames that are similar enough to it are to be extracted for local distance calculation.

A very efficient frame indexing approach was proposed for this purpose [229], [230] by applying locality sensitive hashing techniques on the frames [231], which was shown to be a good approximation for the cosine similarity. Using randomly generated hyperplanes, the posteriorgram space is decomposed into many cone-like regions. These hyperplanes serve as hashing functions, mapping posteriorgrams to one of its sides. For example, by using 64 random hyperplanes, posteriorgrams are transformed into 64 bit values, each bit corresponding to the sides of the hyperplane (the bit value is 1 if the posteriorgram is on one side of the hyperplane, and 0 if it is on the other side). A much simpler approximation for inner product can then be performed by the exclusive-or operation instead of the hamming weight calculation. The posteriorgrams in the documents are therefore sorted by the integer values of their hash values. When searching for document frames similar to a query frame, document frames with integer values within a predefined radius is returned; thus the higher bits are assured identical to the query frame's hash value, whereas lower bits may differ. Since all bits are equally important, several permutations of hash values were performed and sorted; all document frames obtained with each of these permutations are returned if the value is within the radius. This provides a fast filtering to reduce the search space from the whole target content to a limited set of hypothesis frames. Experiments showed that a factor of more than three thousands of real time speedup was achieved by this approach.

4) Information Retrieval based DTW (IR-DTW): This approach [232] was proposed to further speed up the DTW process after the indexed frames in the documents in the target archive were retrieved by the hashing techniques as described above. Instead of going through all points on the (i, j)-plane to check whether a document frame should be retrieved, a vector of retrieved document frames and a vector of extendable path end locations were recorded. In this way the complexity is no longer proportional to the total length of the target archive, but limited by the number of frames returned by the frame indexing approach. By applying path constraints similar to the conventional DTW, and using the frame matching count as a simple measure to estimate the path distance, hypotheses similar to the query can be identified.

D. Modeling Acoustic Patterns for Model-based Approaches

Matching the speech frames with DTW-based approaches is precise and effective, but without ASR much of the underlying linguistic information has been overlooked in the matching process. For example, the speech signals for the same word but produced by different speakers may be very different, as a result the DTW-based approaches may not be able to identify they are referring to the same word, although this is easy with ASR if the recognition is correct.

The above problem comes from the fact that the acoustic characteristics of the speech signals for the same phoneme may vary significantly. In ASR, we use Gaussian mixture models (GMM) or deep neural network (DNN) to model the variations or distributions of such acoustic characteristics based on states in HMMs. The warping function in DTW effectively plays the role of state transitions in HMMs to some extent, but the GMM/DNN modeling of the acoustic characteristic distributions in ASR is actually missing in DTW-based approaches. The posteriorgrams obtained with either GMM or DNN certainly represent ways to take care of the roles of GMM/DNN, although these posteriorgrams are

generated primarily in an unsupervised way and are thus less precise.

On the other hand, speech signals are made of patterns much longer than frames, and the repetitions of similar patterns form the concept of phonemes, syllables and other phonological units. Higher level linguistic units such as words or phrases are then composed of such low level phonological units, and it is these higher level linguistic units which carry semantic information, including the queries we consider here. With a highly effective ASR, speech signals are transcribed into meaningful lexical units such as words, although with recognition errors. When ASR is not performed here with the various considerations mentioned above, it is still possible to learn similar concepts and approaches from ASR, i.e., to train acoustic models to describe the variations or distributions of the acoustic characteristics for some fundamental units in speech signals. The huge target spoken archive can serve as the natural training data for such models, but the difference is that there is no human annotation for the target spoken archive, or the models have to be trained in an unsupervised way. This is reasonable nowadays because huge quantities of spoken archives are available everywhere, but it is very difficult to have human annotation for them.

This leads to the second category of approaches considered here: model-based approaches. Without human annotation, we do not have phonetic knowledge of the audio data any more, but we can identify similar signal patterns having similar acoustic characteristics, referred to as "acoustic patterns" here. Hence, the purpose is to automatically discover the set of acoustic patterns describing the target archive, and train acoustic models for them using the data in the target archive. The approaches here are based on a set of such models trained in an unsupervised way without human annotation. For retrieval purposes, these acoustic patterns should cover the entire target archive, and it is desired that these acoustic patterns can be consistent to some underlying linguistic units such as phonemes. These goals are difficult to achieve, but important along this direction. In this subsection we will first very briefly review some popular approaches for unsupervised discovery of acoustic patterns from an audio data set (the target archive), and training models for these patterns. Use of these models in spoken content retrieval is then presented in the next subsection.

1) Bottom-up modeling: Most approaches for modeling the acoustic patterns follow a three-phase recursive procedure including signal segmentation, unit clustering and model training in each iteration [85], [202], [209], [216], [233]–[239]. In other words, the signals in the archive are first segmented into small units, the units are then clustered into groups based on their acoustic characteristics, and pattern models are finally trained for each group. This process can then be repeated iteratively. A unified nonparametric Bayesian model was developed for jointly modeling the above three subproblems together [217]. In this model, each pattern model is an HMM, and the segment boundaries and the pattern each segment belongs to are hidden variables. This model tries to find the HMM parameters and the hidden variables best representing the audio data collection jointly. These automatically discovered patterns represent phoneme-like (or subword-like) patterns on the highest level in most cases. The above approaches were then extended to include higher level units during training [240], for example, word-like patterns were discovered by identifying the subword-like patterns frequently appearing together. In this way, a lexicon of word-like patterns can be learned and an n-gram language model can be trained on top of these word-like patterns. Semantics were then more or less revealed with these word-like patterns. Experimental results indicated that subword-like patterns generated in this way had high correlation with phoneme identities.

All of the above approaches generate the models bottom-up. Although these approaches modeled the acoustic behaviour of the target spoken archive reasonably well, in most cases they tend to over-cluster the different realizations of the same phonetic identity, e.g., multiple models were very often generated for the same linguistic units such as phonemes. This is reasonable because different realizations for the same phoneme may behave very differently acoustically when produced by different speakers, in different contexts, or under different acoustic conditions. Without human annotation, there is no way to indicate they belong to the same phoneme, and as a result the machine clusters them as different patterns.

For the task of spoken content retrieval, good acoustic patterns need to have high coverage over almost all realizations of the same linguistic identity such as a phoneme. This means the ability for such patterns to model sound characteristics under various conditions is almost indispensable. For example, the realizations of the same vowel produced by male and female speakers are very often split into different acoustic patterns when discovered without human annotation. Without knowing these different patterns refer to the same vowel, we may be able to find only those terms spoken by female speakers when searching with a female spoken query. This is a very challenging problem for approaches along this direction.

2) Top-down Constraints: It has been observed that word-level patterns are easier to identify across speakers than phoneme-level ones [241]. The similarity between the realizations of the same phoneme but produced by different speakers is usually relatively hard to identify, but on the word level, the similarities are very often much more striking. For example, we can usually observe similar formant contours, and similar temporal alternation between voiced/unvoiced segments and low/high frequency energy parts.

With the above observation, a new strategy that tempers the subword-like pattern models obtained from bottom-up training with top-down constraints from the word level was proposed [241]. The repeated word-level patterns are first discovered from the spoken content using techniques such as segmental DTW mentioned in Subsection V-B. For the realizations of the same word-level pattern, DTW alignment between them is then performed. Because they probably have the same underlying subword unit sequences, the DTW aligned acoustic features should therefore map to the same subword units even though they are not acoustically similar. This approach was tested on a task defined earlier [221] different from STD (given a pair of audio segments, the system determined whether they belonged to the same words), but not for STD yet. It was found

that the top-down constraints were capable of improving the performance by up to 57% relative over the bottom-up training alone [241].

3) Transfer Learning: Practically, acoustic patterns do not have to be discovered from scratch. Because all languages are uttered by human beings with a similar vocal tract structure and thereby share some common acoustic patterns, the knowledge obtained from one language can be transferred onto other languages. For resource-rich languages like English, because a huge amount of audio has been collected and annotated, high quality acoustic models are available, and the phonemes represented by these models are known. To transfer the knowledge from a resource-rich language, the target audio (probably in a different language) is decoded by the recognizer of the resource-rich language into phonemes of the resource-rich language, which can be directly used as acoustic patterns in the following spoken content retrieval task, or taken as the initial models for the bottom-up modeling approach [210]. Since the acoustic patterns for one language usually cannot be completely covered by the phoneme set for another and the target audio may include more than one languages, transfer learning from several resource-rich languages, or decoding the target audio with recognizers of several different languages, was shown to be very helpful [210], [211], [242]–[247].

E. Model-based Approaches in Spoken Content Retrieval

With the acoustic patterns discovered and trained from the target spoken archive, different approaches can be applied to perform the model-based signal matching without ASR. Below we present some good examples.

1) Model Posteriorgrams for DTW: A very popular approach is transforming the frame-based acoustic features in both the spoken query and documents into the pattern posteriorgrams, or each signal frame is represented by the posterior probabilities for all acoustic patterns. The DTW-based approaches mentioned in Subsections V-B and V-C can then be directly applied. Experiments on the TIMIT corpus showed that compared to the Gaussian posteriorgrams [209] and RBM posteriorgrams [218], the pattern posteriorgrams from the nonparametric Bayesian model mentioned in Subsection V-D relatively improved the precision by at least 22.1% [217]. It was also shown that the posteriorgrams for the unsupervised acoustic patterns even outperformed the phone posteriorgrams derived from supervised phoneme recognizers if the latter were trained with corpora not matched to the target audio [216], [217], [248].

2) Matching the Query Frames with the Acoustic Pattern Models for the Archive: With a complete set of subword-like patterns, a lexicon of word-like patterns, and a language model for word-like patterns [240], it is possible to decode the target spoken archive off-line into word-like patterns composed of subword-like patterns. The decoding is in exactly the same way as the conventional ASR, but completely unsupervised, with output being the word-like acoustic pattern sequences.

During retrieval, given a spoken query, each frame of acoustic features in the spoken query is matched to the pattern model sequences of the spoken documents in the archive, or evaluated against the HMM states in the pattern models for the documents, very similar to the conventional ASR decoding for which each frame of the input speech is evaluated against the HMM states of the acoustic models [249]. When matching the frame-based query features with the pattern models, a duration-constrained Viterbi algorithm [249] was proposed to avoid unrealistic speaking rate distortion through the matching process, very similar to the slope-constrained DTW discussed earlier in Subsection V-B, except for model-based approach here. The spoken documents are then ranked based on the likelihoods obtained with the Viterbi decoding.

Matching the signal frames in the spoken query with the pattern models representing the target archive actually requires much less computation as compared to the DTW-based approaches, which matches the signal frames in the query with the signal frames in the target archive as mentioned in Subsections V-B and V-C. This is because the numbers of signal frames in the target archive can be huge, but the number of acoustic patterns in the archive can be much less. Experimental results showed a roughly 50% reduction in computation time needed and 2.7% absolute MAP improvement as compared to the segmental DTW approach in Subsection V-B on a Mandarin broadcast news corpus [250].

3) Query Modeling by Pseudo Relevance Feedback: The spoken query can also be represented by pattern models. However, the acoustic patterns are discovered from the archive and therefore can be slightly far from the query. One way to take care of this problem is to train special models (and anti-models) for the query, instead of using the pattern models discovered from the spoken archive. This can be achieved by the pseudo-relevance feedback (PRF) approach introduced in Subsection IV-C [249], [251]. In this approach, a list of hypothesized regions for the spoken query is first generated in the first-pass retrieval, which can be achieved with any unsupervised approach introduced in this section, either DTW-based, or model-based. The top several hypothesized regions on this list that are most possible to be the query are regarded as pseudo-positive examples, while the hypothesized regions that have the lowest confidence scores on the list are regarded as pseudo-negative examples. The pseudo-positive and -negative examples are then used to train respectively a query model and an anti-query model online for exactly the specific query. The final confidence scores of all hypothesized regions on the list are then the likelihood ratio evaluated with the query model and anti-query model for the query.

With this approach, context dependencies among acoustic events inside the queries are better characterized with the query model, while minor signal differences that distinguish the true hypotheses from the false alarms are emphasized by the likelihood ratio. Experimental results showed that this approach offered improved performance if applied on top of either DTW-based or model-based approaches on the TIMIT corpus, Mandarin broadcast news and MediaEval 2011 Spoken Web Search corpus [249], [251].

4) Multi-level Pattern to Pattern Matching across varying Model Configurations: Both the spoken queries and documents can be decoded using the acoustic patterns automatically discovered from the archive, and represented as acoustic pattern sequences. In this way, the matching between the

query and the documents is reduced to comparing the acoustic pattern indices in the pattern sequences, and the on-line computation load can be further reduced because the efficient indexing methods for text content like inverted indexing [252] or WFST-based indexing [66] can be applied. In addition, it was proposed in a recent work that the multi-level sets of acoustic patterns based on varying HMM model granularities (number of states per subword-like pattern model or temporal granularity m, number of distinct subword-like patterns or phonetic granularity n) are complementary to one another, thus can jointly capture the various signal characteristics [253]. It was shown that performing the matching simultaneously over many multi-level sets of patterns is easy, and the integrated scores can offer significantly better performance. This is presented in more details below.

Let $\{p_r, r = 1, 2, 3, .., n\}$ denote the n subword-like patterns in a pattern set. A similarity matrix S of size $n \times n$ is first constructed off-line, for which the element $S(i, j)$ is the similarity between any two pattern HMMs p_i and p_j in the set.

$$S(i, j) = \exp(-\text{KL}(i, j)/\beta), \tag{15}$$

where $\text{KL}(i, j)$ is the KL-divergence between the two pattern HMMs evaluated with the states and summed over the states.

In the on-line phase, the following procedure is performed for the entered spoken query Q and each document D in the archive for each pattern set. Assume for a given pattern set a document D is decoded into a sequence of $|D|$ patterns with indices $(d_1, d_2, ..., d_{|D|})$ and the query Q into a sequence of $|Q|$ patterns with indices $(q_1, ..., q_{|Q|})$. A matching matrix W of size $|D| \times |Q|$ for every document-query pair is thus constructed, in which each entry (i, j) is the similarity between acoustic patterns with indices d_i and q_j as in (16) and shown in Fig 8 for a simple example of $|Q| = 5$ and $|D| = 10$, where the element $S(i, j)$ is defined in (15),

$$W(i, j) = S(d_i, q_j). \tag{16}$$

It is also possible to consider the N-best pattern sequences rather than only the one-best sequence here [253].

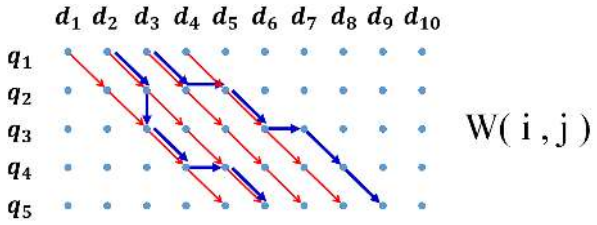

Fig. 8: The matching matrix W in (16) for $D = (d_1, d_2, ..., d_{10})$ and $Q = (q_1, q_2, ..., q_5)$ with subsequence matching (red) or DTW (blue and thicker).

For matching the sub-sequences of D with Q, the elements in the matrix W in (16) are summed along the diagonal direction, generating the accumulated similarities for all subsequences starting at all pattern positions in D as shown in Fig. 8 (red). The maximum is selected to represent the relevance between document D and query Q on the pattern set as in (17).

$$R(Q, D) = \max_i \sum_{j=1}^{|Q|} W(i + j, j). \tag{17}$$

It is also possible to consider dynamic time warping (DTW) on the matrix W as also shown in Fig. 8 (blue and thicker) [253].

The relevance scores $R(Q, D)$ in (17) obtained with all pattern sets based on different model granularities are then averaged, and the average scores are used in ranking all the documents for retrieval. The experiments performed on the TIMIT corpus showed that by integrating the scores obtained with 20 sets of subword-like patterns ($n = 50, 100, 200, 300$ distinct subword-like patterns, $m = 3, 5, 7, 9, 11$ states per pattern HMM), this approach significantly outperformed the DTW-based approach in Subsection V-B by 16.16% in terms of MAP at reduced online computation requirements [253].

VI. Semantic Retrieval of Spoken Content

In this section, we present the next major direction: semantic retrieval of spoken content.

A. Motivation and Background

Most techniques presented above are primarily for STD. Here we shift the focus to semantic retrieval of spoken content. Semantic retrieval has long been highly desired, for which all objects relevant to the query should be retrieved, regardless of including the query terms or not. For example, for a query of "White House", all utterances regarding to the president of United States should be retrieved, although many of them do not include the query "White House". This problem has been widely studied in text information retrieval with many very useful approaches available. Taking the one-best transcriptions from the ASR module as the text, all those technique developed for text information retrieval can be directly applied to semantic retrieval of spoken content, but the ASR errors may seriously degrade the performance. Therefore, special techniques for semantic retrieval of spoken content are necessary. Most of these techniques borrowed some concepts from text information retrieval, but considering the special problems with spoken content. Below, we first very briefly introduce the basic concepts of some techniques for text information retrieval which are useful for spoken content, although much complete information should be found elsewhere [252], [254]. The way to adopt techniques for text retrieval under the framework of cascading speech recognition with text retrieval will then be described. The techniques beyond the cascading framework then follow.

B. Basic Concepts in Text Information Retrieval useful for Semantic Retrieval of Spoken Content

The basic vector space model and language modeling retrieval approach described below provide very good frameworks on top of which query/document expansion techniques can be applied. These techniques were designed for text retrieval, but equally applied for spoken content.

1) Vector Space Model [255]: In this model, documents D and queries Q are respectively represented as vectors $\vec{D}$ and $\vec{Q}$. When the user enters a query Q, the documents D are ranked according to the relevance scores $R(Q, D)$, which is the cosine similarity between $\vec{D}$ and $\vec{Q}$. Each component of $\vec{D}$ and $\vec{Q}$ corresponds to a term t. Typically in text information retrieval, the terms t can be single words, keywords or longer phrases, while for spoken content , subword units or subword unit n-grams are widely considered in addition in order to alleviate the OOV problem. Very often the information based on words and subword units are complementary. The values of the components in the vectors $\vec{D}$ and $\vec{Q}$ corresponding to a term t is represented as $w(t, D)$ and $w(t, Q)$ below. Although there exist different ways to define $w(t, D)$ and $w(t, Q)$, TF-IDF weighting or its variants is the most popularly used. In this weighting scheme, $w(t, D)$ is defined as (18) and (19)..

$$w(t, D) = N(t, D) \times IDF(t), \tag{18}$$

$$IDF(t) = log(T_D/df(t)), \tag{19}$$

where $N(t, D)$ is the total occurrence count for the term t in the document D, or term frequency (TF), and $IDF(t)$ is the inverse document frequency (IDF). T_D is the total number of documents D in the target database, and $df(t)$ is the number of documents containing the term t in the target database. $IDF(t)$ emphasizes those terms t appearing in only very few documents, because they are usually more informative. The definition of $w(t, Q)$ is parallel to $w(t, D)$, except that D in (18) should be replaced by Q.

A major issue in semantic retrieval is that many documents relevant to the query do not necessarily contain the query terms. The IDF here is helpful in this issue. For example, consider the user enters a query "Information Retrieval", which includes two terms, "Information" and "Retrieval". Some relevant documents may only contain the term "Retrieval" but not the whole query of "Information Retrieval". However, if the IDF of the term "Retrieval" is high because it appears only in very limited number of documents, those documents containing only the term "Retrieval" may still have high relevance scores without "Information". On the other hand, the IDF of the term "Information" may be much lower because this term appears in many other documents, so those documents including the term "Information" but not the term "Retrieval" have much lower relevance scores. In this way, some documents having only parts of the query but semantically related to the query may also be retrieved.

2) Language Modeling Retrieval Approach [256], [257]: 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[11]. The relevance score $R(Q, D)$ used to rank the documents D with respect to the given query Q is then the inverse of the KL-divergence between Θ_Q and Θ_D:

$$R(Q, D) = -\text{KL}(\Theta_Q||\Theta_D). \tag{20}$$

[11]There are works to extend the language model from unigrams to also including n-grams and grammars, but out of the scope here [256].

That is, documents whose unigram language models are similar to the query's unigram language model are more likely to be relevant. A document's unigram language model Θ_D is estimated based on the terms in document D as in (21) below.

$$P(t|\Theta_D) = N(t, D)/\sum_t N(t, D), \tag{21}$$

where $N(t, D)$ is as in (18), and Θ_D is usually further interpolated with a background model for smoothing before being used in (20). It has been shown that such smoothing strategies implicitly give higher weights to those rare but informative terms very similar to the inverse document frequency in (19) [258], which is helpful for semantic retrieval. Θ_Q for the query Q is parallel to (21), except that D in (21) is replaced with Q.

3) Query/Document Expansion: Query and document expansion are usually applied to address the problem that all terms in the query are not in the relevant documents, for example, the query is "airplane", whereas there is only "aircraft" in the relevant documents. For document expansion, with latent topic analysis approaches [259]–[262] such as probabilistic latent semantic analysis (PLSA) [260] and latent Dirichlet allocation (LDA) [261], each document vector or document language model can be expanded by assigning non-zero weights in (18) or non-zero probabilities in (21) to those terms not appearing in the document but found semantically related to its content [263]–[266], e.g. adding the term "airplane" to those documents have "aircraft" only, based on the information that the terms "airplane" and "aircraft" may appear in very similar topics. Query expansion can be achieved in similar ways by latent topic analysis, but it was found empirically not as effective as document expansion [267], probably because the queries are usually too short to reliably estimate its latent topics. More effective query expansion is very often realized with pseudo-relevance feedback (PRF) mentioned in Subsection IV-C, i.e., those words appear repeatedly in the documents retrieved in the first pass with the highest scores, but much less frequently in other documents in the target database, can be properly considered and added to the query [167], [169], [268]–[272]. The above document and query expansion techniques developed for text information retrieval can be directly applied on the transcriptions of the spoken content as well [263], [273], [274]. For spoken content retrieval, external information from the web was also shown to be helpful for the expansion of both documents and queries to mitigate the effects of unavoidable ASR errors [275]–[278].

The vector space model and language modeling retrieval approach provide very good frameworks on top of which query and document expansion techniques can be applied in addition. For vector space model, query expansion can be achieved by adding to the original query vector $\vec{Q}$ with the average of the document vectors for the pseudo-relevant documents, and subtracting the average of the vectors for all documents in the database excluding the pseudo-relevant ones [272], so as to add to the query the words appearing repeatedly in the pseudo-relevant documents, but remove from the query those frequently appearing in other documents. For the language modeling retrieval approach, the query expansion can be

formulated by component mixture models [270]. The language models for the pseudo-relevant documents are assumed to be the interpolation of a language model primarily for the query-related terms and a background model for general terms, with document-dependent interpolation weights between the two (e.g. if an irrelevant document is taken as pseudo-relevant, this document's weight for the model for query-related terms should be very low). These document-dependent weights and the two component mixture models are unknown, but can be estimated from the term distributions in the pseudo-relevant documents. Given the estimation, the language model for query-related terms serves as the new query model and is used to replace Θ_Q in (20). In addition, regularizing the estimation process by the original query language model was shown to yield better retrieval performance, and this approach is known as the query-regularized mixture model [167], [259], [260], [263], [265], [266].

C. Estimating TF/IDF Parameters over Lattices

Because the techniques mentioned in Subsection VI-B above were developed for text without errors, the ASR errors may seriously degrade the performance. If the term frequencies $N(t, D)$ in (18) and (21) or inverse document frequencies $IDF(t)$ in (19) are directly counted from the one-best transcriptions, they can be very different from the true values in the spoken content. Therefore, better estimation of these parameters from lattices is crucial. Because the query/document expansion techniques work on top of the vector space model or the language modeling retrieval approach, better TF/IDF parameters are expected to offer better performance.

The expected term frequencies $E(t, D)$ estimated from the lattices are widely used to replace the original term frequencies $N(t, D)$ when applying the vector space model in (18) and language modeling approach in (21) [71], [279], [280].

$$E(t, D) = \sum_{s \in \mathcal{L}(D)} N(t, s) P(s|D), \tag{22}$$

which is parallel to (6), except that the query Q and the utterance u in (6) are respectively replaced by the term t and the spoken document D. By replacing $N(t, D)$ with $E(t, D)$, the vector space model and the language modeling retrieval approach can be very well enhanced [279], [281].

Inverse document frequency for a term t, $IDF(t)$ in (19), is another important parameter for not only the vector space model here, but also many other applications such as summarization and key term extraction. According to (19), inverse document frequency is defined based on $df(t)$, the number of documents in the target database that mention the term t. However, there actually does not exist a well-known good way to estimate this number $df(t)$ from lattices [279][12].

One way to compute $df(t)$ in (19) is to define it to be $\sum_D E(t, D)$ using (22) [82]. However, $IDF(t)$ obtained in this way is certainly quite different from the original idea of inverse document frequency. Another way to obtain $df(t)$ is to take those documents D with expected frequencies of t, $E(t, D)$ in (22), exceeding a threshold as containing t, but there seems to be no good principle in selecting this threshold [71]. There was still another relatively sophisticated approach, in which $df(t)$ is modeled as a linear combination of more than a hundred cues with weights learned from training data [282]. This approach was compared with $df(t)$ estimated on one-best transcriptions or obtained from $E(t, D)$ with a heuristically set threshold, and was shown to yield better retrieval performance based on vector space model [282].

[12]Obviously, it is not a good idea to consider a spoken document with the term t in the lattices as truly containing the term t.

D. Better Estimation of Term Frequencies beyond Directly Averaging over the Lattices

$E(t, D)$ estimated in (22) inevitably suffers from the recognition errors with performance depending on the quality of the lattices. Therefore, some techniques for better calibrating $E(t, D)$ beyond directly averaging over the lattices have been proposed and were shown to offer better results.

For one example, the values of $E(t, D)$ can be modeled as the weighted sum of the scores based on a set of cues obtained from the lattices. With the weights for the cues learned from the training data, better $E(t, D)$ closer to the true frequency count than (22) was shown to be obtainable [283]. Another example is based on the context consistency of the term considered. Because the same term usually have similar context, while quite different context usually implies the terms are different [152]. Therefore, whether a term t exists in a spoken document D can be judged by not only the scores of the arcs hypothesized to be t, but also the word hypotheses of the arcs surrounding the term t in the lattices of D. With some documents containing and not containing the term t as positive and negative examples, a support vector machine (SVM) can be learned to discriminate whether a spoken document truly contains the term t based on the context of t. Then $E(t, D)$ can be better calibrated by decreasing the value if the document D is regarded as not containing t by the SVM and vice versa. Although this approach needs the training data for all the terms t to train an SVM for every term t considered, the training data needed can actually be obtained by pseudo-relevance feedback (PRF) [284] mentioned in Subsection IV-C in practice. $E(t, D)$ calibrated in this way was shown to be able to enhance the document representation in the language modeling retrieval approach, based on which better performance with query expansion was obtained [284].

It is also possible to incorporate some information lost during ASR to better estimate $E(t, D)$ than that in (22) using approaches found useful in Section IV, for example, the graph-based approach solved with random walk as in Subsection IV-E [281], [285]. In this approach, all the arc sequences a whose hypotheses are a specific term t in the lattices obtained from all spoken documents in the whole target archive are clustered into groups based on their time spans, such that those with time spans highly overlapped are in the same group. Each group is represented as a node in a graph for the term t as in Fig. 6, and the edge weights between two nodes are based on the acoustic similarities evaluated with DTW distances between all pairs of acoustic vector sequences corresponding to two arc sequences respectively belonging to the two groups. The initial score of each node is the summation of the posterior probabilities of all its elements. The random

walk algorithm is then performed, and the scores propagated. The new scores for all the groups in the spoken document D are summed over to form a new estimation of the term frequency $E^g(t, D)$ to replace $E(t, D)$ in (22). The above graph construction and random walk are repeated for all t (such as all the words in the lexicon). Different from $E(t, D)$ in (22) which only considers the information from a single lattice, here the acoustic similarity among all arc sequences a whose hypotheses are the considered term t in the lattices of all documents in the entire archive is considered. Experiments performed on Mandarin broadcast news showed that better retrieval performance using document expansion with latent topic analysis and using query expansion with the query-regularized mixture model was achieved [281], no matter the terms t are words, subword units, or segments of several consecutive words or subword units [281].

E. Query Expansion with Acoustic Patterns

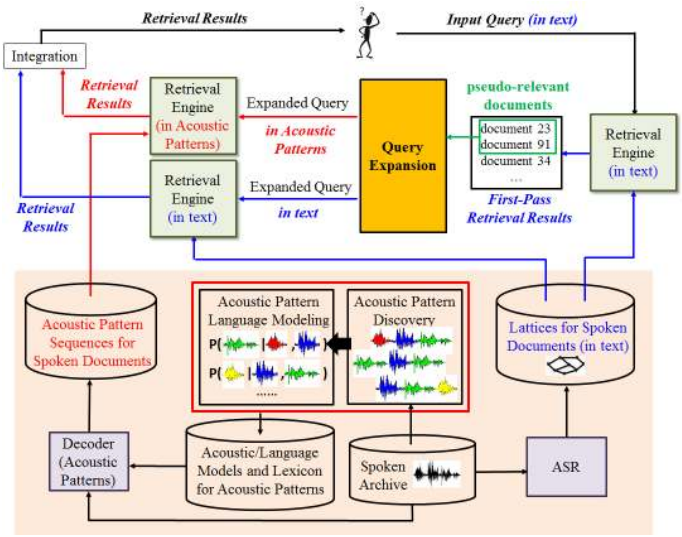

Fig. 9: *The framework of query expansion utilizing automatically discovered acoustic patterns.*

For spoken content retrieval, even if the pseudo-relevant spoken documents actually contain some terms suitable for query expansion[13], these terms may be OOV or incorrectly recognized, never included in the transcriptions or lattices, and therefore cannot help in query expansion. Subword-based query expansion, in which suitable subword sequences are found in the subword-based lattices for query expansion, can address this problem to some extent [274], [287]–[290]. However,the subword-based lattices may have too many incorrect hypotheses, so the subword sequences corresponding to suitable terms for query expansion may not be easy to find.

A new framework of query expansion for semantic retrieval of spoken content was proposed as shown in Fig. 9, in which a set of acoustic patterns automatically learned from the target spoken archive in an unsupervised way as discussed in Subsection V-D is utilized, with a goal to take care of the problem mentioned above [291]. In this work, there are two levels of acoustic patterns used, the word-like patterns, subword-like patterns, plus the lexicon and the language model for the word-like patterns as mentioned in Subsection V-D, all learned from the target spoken archive [240] (lower middle of Fig. 9). As shown of the lower half of Fig. 9, each spoken document is represented in two different forms: lattices in text form (hypothesis of each arc is a word or a subword unit) generated by the conventional ASR module (bottom right corner of Fig. 9), and the one-best acoustic pattern sequences for each spoken document generated by a decoder very similar to ASR module, except based on a set of acoustic/language models and a lexicon for the automatically discovered acoustic patterns [292] (bottom left corner).

When a *text* query is entered, the conventional retrieval engine (upper right of Fig. 9) matches the query terms with the lattices for spoken documents (in text form) to generate the first-pass retrieval results[14]. The top-ranked documents are selected as pseudo-relevant documents. The system then extracts the text terms possibly related to the query from these pseudo-relevant documents to generate the expanded query in text form (upper middle of Fig. 9), which gives a new set of retrieval results via the retrieval engine in text (upper left).

In addition, we have the second version of the expanded query based on acoustic patterns. The acoustic patterns (word-level or subword-level) repeatedly occurring in the pseudo-relevant documents, probably corresponding to some query-related terms but being OOV or incorrectly recognized therefore not present in the lattices obtained with ASR, are also used to form the second expanded query composed of acoustic patterns. Then the expanded query in acoustic patterns is used to retrieve the spoken documents expressed in one-best acoustic pattern sequences. In this way, the acoustic patterns corresponding to some important query-related terms which are OOV or incorrectly recognized by the conventional ASR can be included in the expanded query, and the spoken documents containing these acoustic patterns can thus be retrieved. The results for the two expanded queries are finally integrated (upper left of Fig. 9) and shown to the user. Preliminary experiments on broadcast news showed that the extra query expansion based on acoustic patterns could offer extra improvements than the conventional query expansion based on only the lattices in text form [291].

F. Semantic Retrieval without ASR

Almost all approaches mentioned in Section V achieved without ASR focused on the task of STD by matching the signals directly on the acoustic level without knowing which words are spoken. It seems all they can do is STD. Intuitively semantic retrieval is difficult to achieve without knowing the words, because the semantics or semantic relationships between utterances are carried by or based on words. In experiments on Mandarin Broadcast News [293], the DTW-based query-by-example approach mentioned in

[13]There were also interesting works for "query expansion" for STD, however not for semantic retrieval purpose [286], but to expand the query with the terms phonetically similar to the query. Here we refer to expanding the queries with semantically related but phonetically different terms.

[14] Because the acoustic patterns are discovered in an unsupervised way, the system never knows which text term or which phoneme an acoustic pattern corresponds to. But the query is in text, so the acoustic patterns cannot be used in the first-pass retrieval.

Subsection V-B yielded an MAP score of 28.3% for STD or to return all utterances containing the query terms; but reduced to 8.8% only on the same spoken archive with the same query set using same DTW-based approach when the goal was switched to semantic retrieval, or to return all spoken documents semantically related to the query. This is clearly because many of the spoken documents semantically related to the query didn't contain the query terms, so the DTW-based approaches simply had no way to retrieve these documents. However, some recent work actually managed to achieve the goal of semantic retrieval without ASR to some initial extent as summarized below.

1) Query Expansion without Knowing the Words: When the voice of "United States" is in the original spoken query, we can expand this query with the audio of "America". Then the spoken documents including "America" but not the original query "United States" can also be retrieved. This can be achieved with an ASR module, but becomes difficult without ASR, because the system doesn't know which signal segment corresponds to the words "United States" or "America". Fortunately, the phenomenon that semantically related terms frequently co-occur in the same spoken documents remains true for automatically discovered acoustic patterns with unknown semantics.

We can first use the conventional query-by-example approach (e.g. DTW-based) to locate the documents containing the original spoken query, and then find those acoustic patterns frequently co-occurring with the query in the same documents. Although which words these acoustic patterns correspond to are not known at all, they may correspond to terms semantically related to the original query, so can be added to the original query for expansion. However, the acoustic patterns corresponding to function words usually appear frequently in most spoken documents including those retrieved in the first pass, therefore may also be added to the query and cause interferences. This is why query-regularized mixture model [167] was used to filter out such acoustic patterns for function words [293]. In addition, those spoken documents retrieved by shorter acoustic patterns in the spoken queries should be assigned lower relevance scores [294] because very short acoustic patterns may correspond to subwords rather than real terms. With these approaches, the MAP scores of semantic retrieval without ASR for the experiments on Mandarin broadcast news were improved from 8.8% (DTW-based only) to 9.7% (with query expansion) [293], which was still low, but the improvement was not trivial. This showed that semantic retrieval without ASR is achievable to some extent, although remains to be a very challenging task.

2) Topic Modeling: Topic models learned from the target archive can be helpful for semantic retrieval. The mainstream topic modeling approaches developed for text such as PLSA and LDA can be directly applied on the spoken content when transcribed into text by ASR. This works even with a recognizer for a language different from the target audio. For example, one can transcribe the English audio with a Hungarian phone recognizer, and take the Hungarian phone n-grams as words for topic modeling [295].

Topic modeling can be performed on spoken content even without ASR by taking the automatically discovered acoustic patterns as words. With the topic models, for example, spoken documents can be expanded by acoustic patterns semantically related to its topics but originally not in the documents. The word-level acoustic patterns can also be discovered jointly with the latent topic models [296]. In this approach, segmental DTW mentioned in Subsection V-B was employed first to discover a set of audio intervals, and similar audio intervals very probably sharing the same underlying text transcription were linked together [234]. In this way, the audio intervals and their links actually described the characteristics of the spoken documents without knowing exactly which audio intervals may be instances of which spoken words or phrases. As a result, based on the characteristics of the documents, the acoustic patterns, the probabilities of observing the acoustic patterns given the latent topics, and the latent topic distribution for the spoken documents were jointly learned from the spoken archive. This approach has not yet been applied on semantic retrieval without ASR at the time of writing this article, but the experiments conducted on a set of telephone calls from the Fisher Corpus have demonstrated that the framework successfully provided a means of summarizing the topical structure of an spoken archive by extracting a small set of audio intervals which are actually instances of representative words or phrases for the discovered latent topics [296].

VII. Interactive Retrieval and Efficient Presentation of Retrieved Objects

In this section, we present the next major direction: interactive retrieval and efficient presentation of retrieved objects.

A. Motivation and Background

Most retrieval processes are completed interactively, even for text retrieval. The system returns list of items found, the user selects some of them, and the system further returns more information and so on. This is because the users usually tend to enter short queries not specific enough to describe what they actually intend to find, so very often a few iterations are needed to locate them. For text information, it is easy to extract some snippets for the items found and list them on the screen, and it is easy for the users to find out the desired items at a glance and click on them. Therefore, interactive retrieval is straightforward.

For the spoken content, however, it is not easy to display the retrieved items on the screen, and it is not easy for the user to browse across them, simply because the items are audio (or plus video) signals which can only be played back sequentially, and it is not easy to verify if they include the desired information without completely going through them. The high uncertainty of ASR make the retrieval much less reliable and the interactive process highly desired, but the difficulties in showing the retrieved objects on the screen and having them browsed by the user may make the interactive process very troublesome and discouraging. For example, the subword-based approaches may lead to relatively low precision for the retrieved items, and the user may find it very boring to spend the time to go through the retrieved objects because many of them are irrelevant. Therefore, interactive

retrieval in a way presenting the retrieved items on screen in a comprehensible interface to enable the user to easily navigate across them is crucial. As discussed below, to automatically extract key terms, summaries and generate titles for spoken documents, to automatically construct semantic structures for the spoken content or the retrieved objects, and to offer interactive retrieval in the form of spoken or multi-modal dialogues are possible solutions to these problems.

There have been extensive research aiming for efficient presentation and easy access of spoken (or multimedia) content developed in the past decades or so, some of which were under the scenario of spoken content retrieval, but not all. A few examples are below. The NewsTuner system [297] analyzed the latent semantics of the news and talk radio programs and suggested programs to the user. The Broadcast News Navigator of MITRE [298] answered questions for the news and offered summaries for the news. The National Taiwan University (NTU) Broadcast News Navigator [299] was able to automatically generate titles and summaries for news stories, and organize the news stories in hierarchical structures labelled by automatically extracted key terms under the scenario of interactive spoken content retrieval [300]. The MIT Lecture Browser [19] automatically segmented, transcribed and indexed course lectures and offered efficient ways to retrieve the audio and video segments of the lectures. The FAU Video Lecture Browser displayed automatically extracted key terms for access of video lectures [301].

National Taiwan University (NTU) Virtual Instructor[15], a course lecture system developed at NTU [20], [302], is a good example for the concepts discussed here. Fig. 10 are the example screenshots for the learner/system interactions with the NTU Virtual Instructor for a course on Speech Processing offered at NTU and supported by the system. In Fig. 10 (a), a learner entered the query "triphone" in the blank at the upper right corner, and the retrieval system found a total of 163 utterances in the course containing the query term "triphone". The learner can click the buttons "Play" and listen to the lectures starting with these utterances, or click the links for the slides for lectures including these utterances, for example the slide for the first item with title "5-7 Classification and ....." (in the green frame), to jump to the complete information for the slides. The automatically extracted key terms for the slides are also listed (in the blue frame for the first item) to help the user understand what each slide is all about. If the learner clicked the link for the slide, he saw the screenshot in Fig. 10 (b), where he not only had the slide as on the right, but found that the lecture for this slide was 10 minutes and 23 seconds long (in the green frame), and he could click the bottom "Play Summary" (with the red edges) to listen to a summary of only 1 minute and 2 seconds long. In addition, the learner saw the relationships between all key terms used in this slide and other key terms used in this course automatically extracted from the lectures (referred to as the key term graph here). The key terms of this slide were in a yellow bar (e.g. "classification and regression trees" on the left of the yellow bar), while those key terms below the yellow bar (e.g. "entropy") were the other key terms used in this course related to the one in the yellow bar. If the learner clicked the key term "entropy", as in Fig. 10 (c), the system then showed all slides in the course including this key term and where the key term appeared the first time as an example learning path recommended. Therefore, the learner can choose to learn more about "entropy" sequentially from the beginning or towards more advanced topics if needed.

[15]http://sppc1.ee.ntu.edu.tw/~loach/lecture_2/

B. Summarization, Title Generation and Key Term Extraction for Spoken Documents

Displaying the automatically extracted summaries, titles or key terms can be a good approach to facilitate the browsing of the spoken content, specially across the retrieved objects as summarized below.

1) Summarization: Spoken document summarization [303] has been extensively investigated since 1990s for various purposes not limited to retrieval. Spoken documents in varying genre and domain were considered, including news [299], [304]–[307], meeting records [308]–[311], lectures [302], [312]–[314] and conversational speech [315]. Extractive summarization is usually considered, for which the summary is a set of utterances, phrases or speaker turns automatically selected from a spoken document. The techniques of extractive spoken document summarization spans widely, and only a few examples are mentioned here. A popularly used unsupervised approach is the Maximum Marginal Relevance (MMR) method [316]. It uses a greedy approach for utterance selection and considers the trade-off between the importance of the utterances and the redundancy among the selected utterances. Various approaches were used to estimate the importance of utterances, for example, topic analysis such as PLSA [317], [318]. Another well-known unsupervised approach is the graph-based approach [192], [319], which analyzes the overall relationship among the utterances of a spoken document with a graph using approaches very similar to those explained in Subsection IV-E. With the availability of a set of training documents and their reference summaries, supervised learning can be used. In such cases, the task of extractive summarization was very often considered as a binary classification problem determining whether to include an utterance in the summary [309], [320]. More sophisticated approaches were proposed recently which enable the models to learn to select directly the best utterance subset from a spoken document to be the summary by considering the document as a whole [310], [321]–[324]. In these latter approaches, the different goals such as including important utterances and minimizing redundancy can be jointly learned [321].

Spoken content retrieval and spoken document summarization share some common ground, since both of them need to identify the important components or parts (e.g. keywords) in spoken content. In the mainstream spoken document summarization techniques, the spoken documents are first transcribed into text by ASR, and approaches like MMR and graph-based methods are applied on the ASR outputs. Considering the ASR errors, multiple recognition hypotheses were used [325], and utterances with lower probabilities of being erroneous are selected (e.g. considering confidence scores) [326]. All these can be regarded as the "cascading framework" of spoken

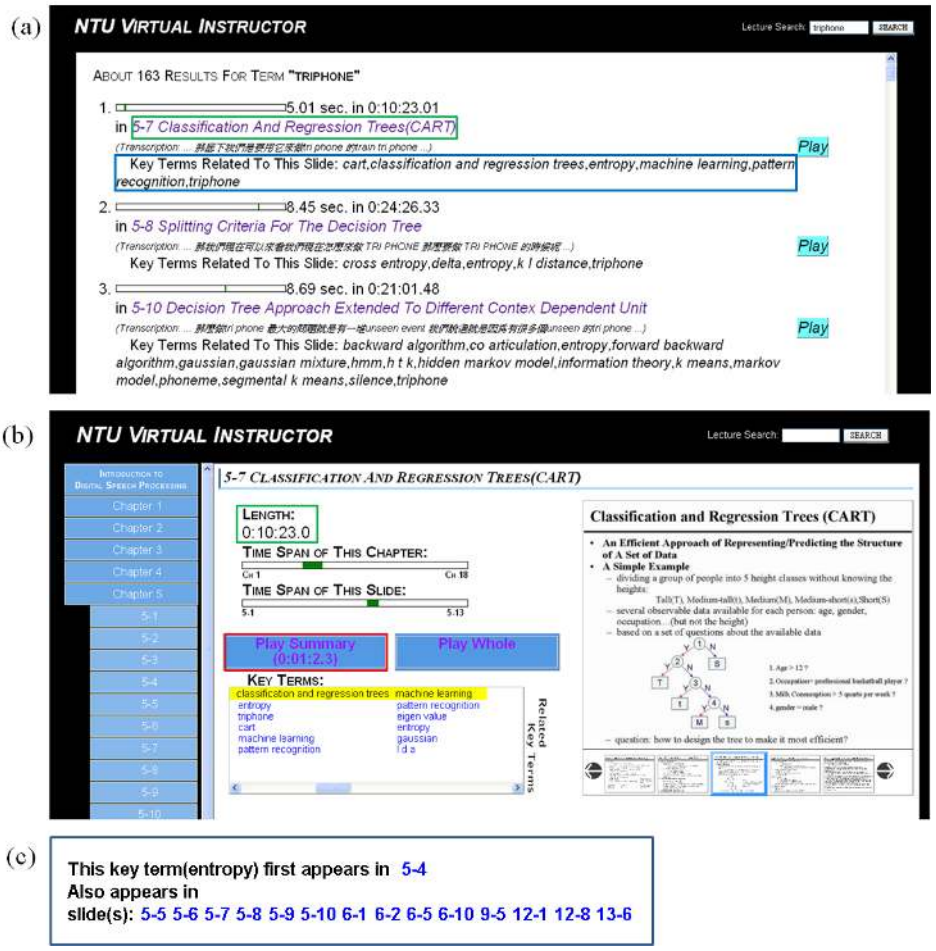

Fig. 10: Example screenshots of NTU Virtual Instructor: (a) spoken content retrieval with input query "triphone", (b) slide, summary and keyterms for the slide with title "5-7 Classification and Regression Tree (CART)" linked from the first item in (a), (c) example learning path for the key term "Entropy" recommended by the system.

document summarization, kind of in parallel to the "cascading framework" of spoken content retrieval. Approaches beyond the "cascading framework" were also proposed. For example, just as ASR can be optimized for spoken content retrieval in Section III, ASR can also be optimized for summarization by considering the word significance in minimum Bayes-risk decoding [327]. In addition, the prosodic features can help not only retrieval as in Section IV-B, but summarization too [321], [328]–[330], since prosodic features help to identify the important part in speech. As spoken content can be retrieved by transfer learning from a different language or even without ASR in Section V, summarizing English spoken documents using a Czech phone recognizer is achievable [331], and by taking the automatically discovered acoustic patterns as words, MMR can also generate good summaries without ASR [332].

2) Title Generation: One example approach is to learn a term selection model, a term ordering model and a title length model from the training corpus including text documents and their human generated titles. The term selection model tells if a term in a document should be selected and used in the title. This includes to select both key terms and the so-called "title terms" (those are not key terms but usually appear in titles). The term ordering model includes strong language models to make sure the order of the selected terms is good and the title is readable. The title length model offers proper length for the title. A Viterbi algorithm is then performed based on the scores from these models over the words used in the summary to generate the title [299], [333].

3) Key Term Extraction: TF-IDF in (18) has been well known to be a good measure for identifying key terms [334], [335], but other measures and approaches beyond the TF-IDF parameters have also been investigated and shown to offer better key terms [336]–[341]. For example, the feature parameters from latent topic models such as PLSA (key terms are usually focused on small number of topics) [339], [341], information from external knowledge resources like Wikipedia [341], and prosodic features extracted from audio signals (key terms are usually produced with slightly lower speed, higher energy and wider pitch range) [340], [341] were found to be useful, and machine learning models were able to provide better solution if some training data with reference key terms were available [336], [340].

C. Semantic Structuring for Spoken Content

This includes global semantic structuring and query-based local semantic structuring as explained below.

Fig. 11: Hierarchical two-dimensional tree structure for global semantic structuring of the spoken content.

1) Global Semantic Structuring: This refers to the task of globally analyzing the semantic structure of the target spoken archive and building the relationships among the individual spoken documents or other kinds of entities such as key terms or named entities. The visualization of the relationships or the structure allows the user to have a convenient and efficient interface to navigate across the target spoken archive. Global semantic structuring has been widely studied for text retrieval and document archives, with WebSOM [342] and ProbMap [343] as good examples, in which the relationships among document clusters are visualized as a two-dimensional map. Another good example is the Google Knowledge Graph [344], which properly connects the entities about people, places and things.

For spoken content, the BBN's Rough'n'Ready system [345] and the Informedia System at Carnegie Mellon University [346] were earlier good examples analyzing the spoken documents in the archive into topics and showing to the user. In the NTU Broadcast News Navigator [347], the spoken documents in the target archive were organized in a hierarchical two-dimensional tree structure for efficient browsing with an example screenshot shown in Fig. 11, in which the clusters of news stories were shown as square blocks on the map and the distances between the blocks reveal the semantic closeness between the clusters. A small set of key terms automatically selected from the news stories in a cluster shown on the block served as the label for that cluster, allowing the user to extract the topics under each cluster. All the clusters in Fig. 11 further belonged to a larger cluster (the block in red) representing a more general concept in another map on the upper layer as shown at the lower left corner of Fig. 11.

In the NTU Virtual Instructor as mentioned above and shown in Fig. 10, a key term graph was constructed from the entire course as a different approach for global semantic structuring [302]. All the key terms automatically extracted from the course were the nodes on the graph, with relationships among the key terms evaluated in different ways based on different features [302]. Only those key terms with high enough relationships in between were connected by edges and shown in the block at the lower left corner of Fig. 10 (b). Each key term was further linked to the lectures for all slides in which the key term was mentioned. Therefore, the lectures for all slides for the entire course were inter-connected through the key terms and the key term graph. In this way, the learner can easily find out related parts of the course and define his own learning path.

2) Query-based Local Semantic Structuring: There were extensive work on local semantic structuring for both text [348] and spoken content [349], [350]. The retrieved objects for such given query are clustered on-line with algorithms such as the Hierarchical Agglomerative Clustering (HAC) to construct a topic hierarchy [349], [350]. Each cluster of semantically similar objects is a node on the hierarchy, and one or few key terms are selected from the cluster to be used as the label for the node. In this way, the user can easily select or delete a node when browsing over the hierarchy.

D. Interaction with Spoken or Multi-modal Dialogues

U1:	US President, please.
S1:	Your query is ambiguous. More precisely, please?
U2:	Diplomatic issue.
S2:	Persian Gulf?
U3:	No.
S3:	Please view the list and select one item relevant to your need.
U4:	(Pick the document at rank 6th, "Obama: We Welcome China's Rise, January 19, 2011 3:47 PM, CBSNEWS")
S4:	(Show the list) Here are what you are looking for.

Fig. 12: An example scenario of interactive spoken content retrieval between the system (**S**) and the user (**U**).

Interactive information retrieval (IIR) has been used for about two decades to make the retrieval process more effective [351]. The Dialogue Navigator for Kyoto City [352] is a very good example, which helps users navigate across Wikipedia documents about Kyoto as well as the tourist information from the Kyoto city government.

Fig. 12 is a possible interaction scenario for retrieving broadcast news stories [353]. Suppose a user is looking for the news about the meeting of US President Barack Obama with the leader of China. He may simply enter the short query of "US President" (U1), which is ambiguous since there are many news stories on completely different topics in the archive related to "US President". The system finds the retrieved objects have topics diverging widely, thus asks the user for further information (S1), and receives the next instruction, "Diplomatic issue"(U2). With this second instruction, the system finds many news items retrieved with the query "US President" plus "Diplomatic issue" have a common key term of Persian Gulf, so the system further clarifies with the user if he wishes to find news related to "Persian Gulf "(S2) and gets the answer "No" (U3). This answer significantly narrows down the target, and therefore the system offers a list of example items for the user to select, very probably each of which represents a somewhat different topic (S3). With the selection of the example spoken document (U4), the system then has enough information to retrieve the documents the user is looking for, so the final retrieval results are presented to the user (S4).

The above interactive process is actually a multi-modal dialogue (spoken dialogue plus other means of interaction). Such dialogue processes have been well studied for other tasks such as air ticket booking, city guides, and so on [352], [354],

[355], so extending experiences in those tasks to interactive retrieval is natural, for example, considering a statistical model such as a Markov Decision Process (MDP) [356]. In MDP, the actions taken by the system is chosen based on the states, which can be one or more continuous or quantized values (here the estimated quality of the present retrieved results based on all the input entered so far by the user (U1,U2,U3,U4)). The system can take different types of actions (e.g. asking for more information (S1), requesting for confirmation with a key term (S2), returning a list of examples for selection (S3), etc.) on different states to clarify the user's intention based on an intrinsic policy. This policy can be optimized based on a pre-defined reward function with reinforcement learning (e.g. the fitted value iteration (FVI) [357] algorithm) using a corpus of historical data of user interactions, or simulated users generated based on some of such data [358]. The state can be estimated based on the values of some performance measures of the retrieval results such as MAP mentioned in Subsection II-C2 [353], [359], while the key terms can be obtained as mentioned in Subsection VII-B3. As a result, the system is able to choose the proper actions to interact with the user at each stage of the retrieval process such that the retrieval performance can be maximized while the extra burden for the user can be minimized.

VIII. Concluding Remarks and Prospect

Many advanced application tasks of spoken language processing were solved by cascading a set of modules in early stages of developments. Take the spoken dialogue system as an example, which was actually built in early years by cascading ASR, natural language understanding, dialogue management, natural language generation and TTS [360]. Today the spoken dialogue is already a full-fledged independent area far beyond the above cascading framework. Good examples include the dialogue managers based on Partially Observable Markov Decision Process (POMDP) taking the uncertainty of ASR and spoken language understanding into considerations [361], and learning the policy of dialogue manager and natural language generator jointly [362]. These novel techniques beyond the cascading framework have turned the pages of the research and development of spoken dialogues. Another example is speech translation, in which jointly optimizing the ASR module and its downstream text processing module is also considered as a major trend [105]. We believe similar developments may be experienced in spoken content retrieval in the future.

Cascading ASR with text retrieval has been very successful in this area, but inevitably becomes less adequate for more challenging real-world tasks. This is why the concepts beyond the cascading framework become important, which is categorized into five major directions as in Sections III, IV, V, VI and VII. Below we make brief concluding remarks for each of them.

(1) Modified Speech Recognition for Retrieval Purposes (Section III): Here the ASR and text retrieval are still cascaded, but ASR is properly modified or learned for retrieval purposes. Quite several approaches here are based on a known query set, therefore limited to the scenario of keyword spotting currently. Hence, one next step is to try to generalize these approaches to unknown queries during training. Relevance feedback in Subsection III-B3 is a good way to collect training data, not only for learning retrieval-oriented acoustic models as mentioned here, but for learning retrieval-oriented language models and ASR output transformation, and it is also possible to replace relevance feedback with PRF in Subsection IV-C. In the long run, a more compact integration of ASR and retrieval may be possible, and an initial version of it may be similar to the one described in Subsection III-F.

(2) Exploiting Information Not Present in Standard ASR Transcriptions (Section IV): The information in speech signals but not present in ASR outputs can be better utilized. Quite several approaches here used query-specific rescoring based on the similarity between the signal segments in the target archive hypothesized as the query. The similarity was usually computed by DTW, but because DTW is limited in considering signal distributions, replacing DTW by model-based approaches in Subsections V-D and V-E could be better. Because rescoring is based on the first-pass results, the performance is limited by the recall of the first pass. Improving the recall by fuzzy matching or subword-based retrieval can make rescoring more powerful [178], [363]. Of course, it would be very attractive if we could use the information in the speech signals directly without relying on the first pass, but no work in this way has been reported yet.

(3) Directly Matching on Acoustic Level without ASR (Section V): For spoken queries, the signals can be directly matched on the acoustic level rather than the phoneme or word levels, so all problems with ASR can be bypassed. This matching can be based on DTW, but the model-based approaches based on the acoustic patterns discovered from the target spoken archive may be better in coping with the signal variations. The achieved performance along this direction is still not comparable with those with ASR. However, with the Big Data generated every day and improved pattern discovery techniques, the performance gap may be narrowed, although there is still a very long way to go.

(4) Semantic Retrieval of Spoken Content (Section VI): Retrieving semantically related spoken content not necessarily including the query is still a very ambitious goal. It didn't attract as much attention as STD maybe because the entry barrier is higher, including the difficulty of annotating semantically related query-document data sets for the experiments, and the annotation may even be subjective. With some benchmark data sets becoming available in recent years, such as the semantic retrieval task of *NTCIR*[16] *SDR* [53] and *Question Answering for Spoken Web* [294], more work can hopefully be developed nowadays.

(5) Interactive Retrieval and Efficient Presentation of Retrieved Objects (Section VII): The spoken content is difficult to be shown on the screen and browsed by the user, so the techniques for efficiently presenting the retrieved objects on an interactive interface are highly desired. Key term extraction, title generation, summarization, and semantic structuring for spoken content are all useful techniques for presenting the spoken content, but they are still very challenging tasks today,

[16]NII Testbeds and Community for Information access Research

and better approaches are yet to be developed. Learning the experiences from text document processing area on these problems may be helpful. Also, much more experiences in human-machine interactions are still to be learned from the very successful discipline of spoken dialogues.

On the other hand, most works along the above five directions were proposed and developed individually. Very wide space for integration among the five directions are actually possible, although very limited results have been reported. Directions 1 (Section III) and 2 (Section IV) are actually orthogonal and can be combined to offer better results. One such example was mentioned at the end of Subsection IV-D2. Direction 3 (Section V) doesn't use ASR so sounds different, but the acoustic patterns in that direction can be used with Direction 2 as mentioned above, hopefully also helpful to Direction 1. Hence, we believe Direction 3 is also orthogonal to Directions 1 and 2. Directions 4 (Section VI) and 5 (Section VII) are orthogonal to each other, and orthogonal to Directions 1, 2 and 3, so they add two extra dimensions. Good examples are in Subsection VI-D (using Direction 2 in Direction 4) and Subsection VI-E and VI-F (using Direction 3 in Direction 4), although Direction 1 seemed not yet used in Direction 4. The five directions open quite wide space for future developments. Of course, we also look forward to seeing extra directions beyond the above five directions we have seen presently.

The success of text content retrieval is a major reason of how the Internet has become an indispensable part of our daily lives. If spoken content retrieval can be successful, our daily lives may be further changed and very different. Consider an example scenario referred to as spoken knowledge organization here [302]. With the necessity of life-long learning in the era of knowledge explosion and the rapid proliferation of Massive Open Online Courses (MOOCs), worldwide instructors are posting slides and video/audio recordings of their lectures on on-line platforms, and worldwide learners can easily access the curricula. However, a major difficulty for the learners is that it may not be easy for them to spend tens of hours to go through a complete course, but the course content is usually sequential. It is not easy to understand a lecture segment without learning the background, but it is even more difficult to find where the necessary background is. Because the speech signals tell exactly the knowledge being conveyed in these lectures, successful spoken content retrieval technologies may be able to locate exactly the parts of the course lectures matched to the learners' needs, as well as the necessary background or relevant information for the required knowledge, all of which may spread over many different courses offered by many different instructors. This may lead to the highly desired personalized learning environment for the large number of worldwide online learners working on different task domains with different background knowledge and widely varying learning requirements.

Another example scenario depicting the way our daily lives may be changed and become very different because of successful spoken content retrieval technologies is referred to as multimedia information management here, or the technologies that can find, filter, select and manage the information the user needs from the heterogeneous multimedia resources over the Internet. Assume a user types a query "David Beckham" (the name of a globally renowned English former footballer), in addition to offering the related web pages as what typical search engines do today, the video recordings of the exciting moments for the historic games David Beckham participated in may also be retrieved from the video sharing platforms based on the audio parts of the videos. The exciting moments in each of these historic games can even be automatically summarized by jointly analyzing the video frames and the audio of the commentators. The interview videos with David Beckham after these games and the videos about the stories of David Beckham's family lives and family members can also be similarly linked. Note that for these videos the key information is actually in the spoken part, so successful spoken content retrieval technologies integrated with other information management technologies may realize the above scenario. However, today the search for such videos still rely on the very limited text descriptions of the videos rather than the spoken content, but only successful spoken content retrieval can locate exactly the desired video frames carrying the desired information. These example scenarios show that successful spoken content retrieval may bring further major changes to our daily lives. We are all working towards that goal, and looking forward to its realization in the future.

REFERENCES

[1] G. Tur and R. DeMori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons Inc, 2011, ch. 15, pp. 417–446.

[2] C. Chelba, T. Hazen, and M. Saraclar, "Retrieval and browsing of spoken content," *Signal Processing Magazine, IEEE*, vol. 25, no. 3, pp. 39 –49, may 2008.

[3] M. Larson and G. J. F. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Found. Trends Inf. Retr.*, vol. 5, pp. 235–422, 2012.

[4] L.-s. Lee and Y.-C. Pan, "Voice-based information retrieval – how far are we from the text-based information retrieval ?" in *ASRU*, 2009.

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

[6] K. Koumpis and S. Renals, "Content-based access to spoken audio," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 61–69, 2005.

[7] A. Mandal, K. Prasanna Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, pp. 1–16, 2013.

[8] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "A lattice-based approach to query-by-example spoken document retrieval," in *SIGIR*, 2008.

[9] M. Saraclar, "Lattice-based search for spoken utterance retrieval," in *HLT-NAACL*, 2004, pp. 129–136.

[10] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 2005, pp. 443–450.

[11] D. Vergyri, I. Shafran, A. Stolcke, R. R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 spoken term detection system," in *Interspeech*, 2007.

[12] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *SIGIR*, 2007.

[13] D. R. H. Miller, M. Kleber, C. lin Kao, O. Kimball, T. Colthurst, S. A. Lowe, R. M. Schwartz, and H. Gish, "Rapid and accurate spoken term detection," in *Interspeech*, 2007.

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

[15] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, *The TREC Spoken Document Retrieval Track: A Success Story*, 2000.

[16] *http://speechfind.utdallas.edu/*.

[17] J. Ogata and M. Goto, "Podcastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription," in *Interspeech*, 2009.

[18] 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.

[19] 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.

[20] 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 and semantic structuring for spoken documents," in *ICASSP*, 2009.

[21] H.-Y. Lee, Y.-L. Tang, H. Tang, and L.-s. Lee, "Spoken term detection from bilingual spontaneous speech using code-switched lattice-based structures for words and subword units," in *ASRU*, 2009.

[22] G. Heigold, H. Ney, R. Schluter, and S. Wiesler, "Discriminative training for automatic speech recognition: Modeling, criteria, optimization, implementation, and performance," *Signal Processing Magazine, IEEE*, vol. 29, pp. 58–69, 2012.

[23] M. Gales, S. Watanabe, and E. Fosler-Lussier, "Structured discriminative models for speech recognition: An overview," *Signal Processing Magazine, IEEE*, vol. 29, pp. 70–81, 2012.

[24] G. Saon and J.-T. Chien, "Large-vocabulary continuous speech recognition systems: A look at some recent advances," *Signal Processing Magazine, IEEE*, vol. 29, pp. 18–33, 2012.

[25] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, pp. 82–97, 2012.

[26] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 30–42, 2012.

[27] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *ICASSP*, 2010.

[28] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, 2003.

[29] S. Huang and S. Renals, "Hierarchical bayesian language models for conversational speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 1941–1954, Nov 2010.

[30] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge University Engineering Dept, 2003.

[31] T. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, D. Van Compernolle, K. Demuynck, J. Gemmeke, J. Bellegarda, and S. Sundaram, "Exemplar-based processing for speech recognition: An overview," *Signal Processing Magazine, IEEE*, vol. 29, pp. 98–113, 2012.

[32] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *Interspeech*, 2010.

[33] R. Stern and N. Morgan, "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, pp. 34–43, 2012.

[34] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, pp. 1060–1089, 2013.

[35] D. Yu, L. Deng, and F. Seide, "The deep tensor neural network with applications to large vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 388–396, 2013.

[36] Y. Kubo, S. Watanabe, T. Hori, and A. Nakamura, "Structural classification methods based on weighted finite-state transducers for automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 2240–2251, 2012.

[37] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 14–22, 2012.

[38] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *ICASSP*, 2014.

[39] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. Sainath, and A. Sethy, "Developing speech recognition systems for corpus indexing under the IARPA babel program," in *ICASSP*, 2013, pp. 6753–6757.

[40] B. Kingsbury, J. Cui, X. Cui, M. Gales, K. Knill, J. Mamou, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. Woodland, "A high-performance cantonese keyword search system," in *ICASSP*, 2013.

[41] J. Mamou, J. Cui, X. Cui, M. Gales, B. Kingsbury, K. Knill, L. Mangu, D. Nolden, M. Picheny, B. Ramabhadran, R. Schluter, A. Sethy, and P. Woodland, "System combination and score normalization for spoken term detection," in *ICASSP*, 2013.

[42] D. Karakos, R. Schwartz, S. Tsakalidis, L. Zhang, S. Ranjan, T. Ng, R. Hsiao, G. Saikumar, I. Bulyko, L. Nguyen, J. Makhoul, F. Grezl, M. Hannemann, M. Karafiat, I. Szoke, K. Vesely, L. Lamel, and V.-B. Le, "Score normalization and system combination for improved keyword spotting," in *ASRU*, 2013.

[43] J. Cui, J. Mamou, B. Kingsbury, and B. Ramabhadran, "Automatic keyword selection for keyword search development and tuning," in *ICASSP*, 2014.

[44] R. Hsiao, T. Ng, F. Grezl, D. Karakos, S. Tsakalidis, L. Nguyen, and R. Schwartz, "Discriminative semi-supervised training for keyword search in low resource languages," in *ASRU*, 2013.

[45] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *Proc. ICASSP*, 2014.

[46] S. Wegmann, A. Faria, A. Janin, K. Riedhammer, and N. Morgan, "The TAO of ATWV: Probing the mysteries of keyword search performance," in *ASRU*, 2013.

[47] A. Gandhe, F. Metze, A. Waibel, and I. Lane, "Optimization of neural network language models for keyword search," in *ICASSP*, 2014.

[48] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic, "Real life information retrieval: a study of user queries on the web," *SIGIR Forum*, vol. 32, pp. 5–17, 1998.

[49] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *Signal Processing Magazine, IEEE*, vol. 25, pp. 28–38, 2008.

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

[51] *http://www.iarpa.gov/Programs/ia/Babel/babel.html*.

[52] *http://www.darpa.mil/Our_Work/I2O/Programs/Robust_Automatic_Transcription_of_Speech_(RATS).aspx*.

[53] T. Akiba, H. Nishizaki, K. Aikawa, T. Kawahara, and T. Matsui, "Overview of the IR for spoken documents task in NTCIR-9 workshop," in *Proceedings of NTCIR-9 Workshop*, 2011.

[54] J. S. Garofolo, E. M. Voorhees, C. G. P. Auzanne, V. M. Stanford, and B. A. Lund, "1998 trec-7 spoken document retrieval track overview and results," in *Proc. 7th Text Retrieval Conference TREC-7*, 1999.

[55] E. M. Voorhees and D. Harman, "Overview of the eighth text retrieval conference (trec-8)," 2000, pp. 1–24.

[56] G. Tur and R. DeMori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons Inc, 2011, ch. 11, pp. 257–280.

[57] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. Cambridge University Press, 2008, ch. 8.

[58] *Text REtrieval Conference, http://trec.nist.gov/*.

[59] P. Yu, K. Chen, L. Lu, and F. Seide, "Searching the audio notebook: keyword search in recorded conversations," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

[60] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," in *ICASSP*, 2008.

[61] A. Norouzian and R. Rose, "An efficient approach for two-stage open vocabulary spoken term detection," in *SLT*, 2010.

[62] K. Iwami, Y. Fujii, K. Yamamoto, and S. Nakagawa, "Efficient out-of-vocabulary term detection by n-gram array indices with distance from a syllable lattice," in *ICASSP*, 2011.

[63] K. Thambiratnam and S. Sridharan, "Rapid yet accurate speech indexing using dynamic match lattice spotting," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2007.

[64] J. Cernocky, I. Szoke, M. Fapso, M. Karafiat, L. Burget, J. Kopecky, F. Grezl, P. Schwarz, O. Glembek, I. Oparin, P. Smrz, and P. Matejka, "Search in speech for public security and defense," in *Signal Processing Applications for Public Security and Forensics*, 2007.

[65] I. Szoke, L. Burget, J. Cernocky, and M. Fapso, "Sub-word modeling of out of vocabulary words in spoken term detection," in *SLT*, 2008.

[66] 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*, 2004.

[67] D. Can and M. Saraclar, "Lattice indexing for spoken term detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 2338–2347, 2011.

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

[69] D. Can, E. Cooper, A. Sethy, C. White, B. Ramabhadran, and M. Saraclar, "Effect of pronunciations on OOV queries in spoken term detection," in *ICASSP*, 2009.

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

[71] J. Mamou, D. Carmel, and R. Hoory, "Spoken document retrieval from call-center conversations," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '06, 2006, pp. 51–58.

[72] J. Silva, C. Chelba, and A. Acero, "Pruning analysis for the position specific posterior lattices for spoken document search," in *ICASSP*, 2006.

[73] ——, "Integration of metadata in spoken document search using position specific posterior latices," in *SLT*, 2006.

[74] Z.-Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken-document retrieval for the internet: lattice indexing for large-scale web-search architectures," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 2006, pp. 415–422.

[75] F. Seide, P. Yu, and Y. Shi, "Towards spoken-document retrieval for the enterprise: Approximate word-lattice indexing with text indexers," in *ASRU*, 2007.

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

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

[78] 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.

[79] 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, no. 5, pp. 899 – 906, oct. 2005.

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

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

[82] V. T. Turunen and M. Kurimo, "Indexing confusion networks for morph-based spoken document retrieval," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007.

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

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

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

[86] C. Dubois and D. Charlet, "Using textual information from LVCSR transcripts for phonetic-based spoken term detection," in *ICASSP*, 2008.

[87] A. Rastrow, A. Sethy, B. Ramabhadran, and F. Jelinek, "Towards using hybrid word and fragment units for vocabulary independent LVCSR systems," in *Interspeech*, 2009.

[88] I. Szoke, M. Fapso, L. Burget, and J. Cernocky, "Hybrid word-subword decoding for spoken term detection," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008.

[89] D. Wang, S. King, and J. Frankel, "Stochastic pronunciation modelling for spoken term detection," in *Interspeech*, 2009.

[90] D. Wang, S. King, N. Evans, and R. Troncy, "CRF-based stochastic pronunciation modeling for Out-of-Vocabulary spoken term detection," in *Interspeech*, 2010.

[91] D. Wang, S. King, J. Frankel, and P. Bell, "Stochastic pronunciation modelling and soft match for Out-of-Vocabulary spoken term detection," in *ICASSP*, 2010.

[92] 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.

[93] Y.-C. Pan and L.-s. Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1562–1574, Aug 2010.

[94] S. Parlak and M. Saraclar, "Performance analysis and improvement of turkish broadcast news retrieval," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 731–741, 2012.

[95] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in *ASRU*, 2013.

[96] D. Can and M. Saraçlar, "Score distribution based term specific thresholding for spoken term detection," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, 2009.

[97] B. Zhang, R. Schwartz, S. Tsakalidis, L. Nguyen, and S. Matsoukas, "White listing and score normalization for keyword spotting of noisy speech," in *Interspeech*, 2012.

[98] "BBN presentation at IARPA babel PI meeting," Sep 2012.

[99] V. T. Pham, H. Xu, N. C. F. Yih, S. Sivadas, L. B. Pang, C. E. Siong, and L. Haizhou, "Discriminative score normalization for keyword search decision," in *ICASSP*, 2014.

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

[101] M. Akbacak, L. Burget, W. Wang, and J. van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *ICASSP*, 2013.

[102] S. Meng, W.-Q. Zhang, and J. Liu, "Combining chinese spoken term detection systems via side-information conditioned linear logistic regression," in *Interspeech*, 2010.

[103] Y.-H. Chen, C.-C. Chou, H.-Y. Lee, and L.-s. Lee, "An initial attempt to improve spoken term detection by learning optimal weights for different indexing features," in *ICASSP*, 2010, pp. 5278 –5281.

[104] C.-H. Meng, H.-Y. Lee, and L.-s. Lee, "Improved lattice-based spoken document retrieval by directly learning from the evaluation measures," in *ICASSP*, 2009.

[105] X. He and L. Deng, "Speech-centric information processing: An optimization-oriented approach," *Proceedings of the IEEE*, vol. 101, pp. 1116–1135, 2013.

[106] ——, "Optimization in speech-centric information processing: Criteria and techniques," in *ICASSP*, 2012.

[107] L. van der Werff and W. Heeren, "Evaluating ASR output for information retrieval," in *SSCS*, 2007.

[108] S. Johnson, P. Jourlin, G. L. Moore, K. S. Jones, and P. Woodland, "The cambridge university spoken document retrieval system," in *ICASSP*, 1999.

[109] M. Larson, M. Tsagkias, J. He, and M. de Rijke, "Investigating the global semantic impact of speech recognition error on spoken content collections," in *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2009.

[110] J. Shao, R.-P. Yu, Q. Zhao, Y. Yan, and F. Seide, "Towards vocabulary-independent speech indexing for large-scale repositories," in *Interspeech*, 2008.

[111] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, "The effect of language models on phonetic decoding for spoken term detection," in *SSCS*, 2009.

[112] J. S. Olsson, "Improved measures for predicting the usefulness of recognition lattices in ranked utterance retrieval," in *SSCS*, 2007.

[113] B.-H. Juang, W. Hou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 3, pp. 257 –265, may 1997.

[114] D. Povey and P. Woodland, "Minimum phone error and i-smoothing for improved discriminative training," in *ICASSP*, 2002.

[115] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Interspeech*, 2013.

[116] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *Signal Processing Magazine, IEEE*, 2008.

[117] Q. Fu and B.-H. Juang, "Automatic speech recognition based on weighted minimum classification error (W-MCE) training method," in *ASRU*, 2007.

[118] C. Weng, B.-H. F. Juang, and D. Povey, "Discriminative training using non-uniform criteria for keyword spotting on spontaneous speech," in *Interspeech*, 2012.

[119] C. Weng and B.-H. F. Juang, "Adaptive boosted non-uniform mce for keyword spotting on spontaneous speech," in *ICASSP*, 2013.

[120] Q. Fu, Y. Zhao, and B.-H. Juang, "Automatic speech recognition based on non-uniform error criteria," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 780–793, 2012.

[121] I.-F. Chen, N. F. Chen, and C.-H. Lee, ",a keyword-boosted sMBR criterion to enhance keyword search performance in deep neural network based acoustic modeling," in *Interspeech*, 2014.

[122] S. Tsakalidis, X. Zhuang, R. Hsiao, S. Wu, P. Natarajan, R. Prasad, and P. Natarajan, "Robust event detection from spoken content in consumer domain videos," in *Interspeech*, 2012.

[123] I.-F. Chen and C.-H. Lee, "A resource-dependent approach to word modeling for keyword spotting," in *Interspeech*, 2013.

[124] ——, "A study on using word-level HMMs to improve ASR performance over state-of-the-art phone-level acoustic modeling for LVCSR," in *Interspeech*, 2012.

[125] A. Jansen and P. Niyogi, "Point process models for spotting keywords in continuous speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, 2009.

[126] K. Kintzley, A. Jansen, and H. Hermansky, "Featherweight phonetic keyword search for conversational speech," in *ICASSP*, 2014.

[127] C. Liu, A. Jansen, G. Chen, K. Kintzley, J. Trmal, and S. Khudanpur, "Low-resource open vocabulary keyword search using point process models," in *Interspeech*, 2014.

[128] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," in *The Knowledge Engineering Review*, 2003, pp. 95–145.

[129] 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, no. 7, pp. 2095 –2110, sept. 2012.

[130] 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.

[131] ——, "A framework integrating different relevance feedback scenarios and approaches for spoken term detection," in *SLT*, 2012.

[132] H.-Y. Lee and L.-s. Lee, "Improving retrieval performance by user feedback: a new framework for spoken term detection," in *ICASSP*, 2010.

[133] J. Thorsten, "Optimizing search engines using clickthrough data," in *KDD*, 2002.

[134] J.-T. Chien and M.-S. Wu, "Minimum rank error language modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, pp. 267 –276, 2009.

[135] A. Mandal, J. V. Hout, Y.-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, and H. Franco, "Strategies for high accuracy keyword detection in noisy channels," in *Interspeech*, 2013.

[136] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen, and C.-H. Lee, "A novel keyword+LVCSR-filler based grammar network representation for spoken keyword search," in *ISCSLP*, 2014.

[137] P. Karanasou, L. Burget, D. Vergyri, M. Akbacak, and A. Mandal, "Discriminatively trained phoneme confusion model for keyword spotting," in *Interspeech*, 2012.

[138] M. Saraclar, A. Sethy, B. Ramabhadran, L. Mangu, J. Cui, X. Cui, B. Kingsbury, and J. Mamou, "An empirical study of confusion modeling in keyword search for low resource languages," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, 2013.

[139] R. Wallace, R. Vogt, B. Baker, and S. Sridharan, "Optimising figure of merit for phonetic spoken term detection," in *ICASSP*, 2010.

[140] R. Wallace, B. Baker, R. Vogt, and S. Sridharan, "Discriminative optimization of the figure of merit for phonetic spoken term detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1677–1687, 2011.

[141] Y. Y. Tomoyosi Akiba, "Spoken document retrieval by translating recognition candidates into correct transcriptions," in *Interspeech*, 2008.

[142] K. Thambiratnam and S. Sridharan, "Dynamic match phone-lattice searches for very fast and accurate unrestricted vocabulary keyword spotting," in *ICASSP*, 2005.

[143] T. Mertens and D. Schneider, "Efficient subword lattice retrieval for german spoken term detection," in *ICASSP*, 2009.

[144] U. V. Chaudhari and M. Picheny, "Improvements in phone based audio search via constrained match with high order confusion estimates," in *ASRU*, 2007.

[145] T. Mertens, D. Schneider, and J. Kohler, "Merging search spaces for subword spoken term detection," in *Interspeech*, 2009.

[146] U. V. Chaudhari and M. Picheny, "Matching criteria for vocabulary-independent search," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, pp. 1633–1643, 2012.

[147] R. Prabhavalkar, K. Livescu, E. Fosler-Lussier, and J. Keshet, "Discriminative articulatory models for spoken term detection in low-resource conversational settings," in *ICASSP*, 2013.

[148] R. Prabhavalkar, J. Keshet, K. Livescu, and E. Fosler-Lussier, "Discriminative spoken term detection with limited data," in *MLSLP*, 2012.

[149] M. Wollmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," in *ICASSP*, 2009.

[150] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317 – 329, 2009.

[151] ——, "Discriminative keyword spotting," in *NOLISP*, 2007.

[152] D. Schneider, T. Mertens, M. Larson, and J. Kohler, "Contextual verification for open vocabulary spoken term detection," in *Interspeech*, 2010.

[153] T.-W. Tu, H.-Y. Lee, and L.-s. Lee, "Improved spoken term detection using support vector machines with acoustic and context features from pseudo-relevance feedback," in *ASRU*, 2011.

[154] H.-Y. Lee, T.-W. Tu, C.-P. Chen, C.-Y. Huang, and L.-s. Lee, "Improved spoken term detection using support vector machines based on lattice context consistency," in *ICASSP*, 2011.

[155] M. Seigel, P. Woodland, and M. Gales, "A confidence-based approach for improving keyword hypothesis scores," in *ICASSP*, 2013.

[156] H. Li, J. Han, T. Zheng, and G. Zheng, "A novel condence measure based on context consistency for spoken term detection," in *Interspeech*, 2012.

[157] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves oov detection in speech," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[158] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," in *ICASSP*, 2010.

[159] J. Tejedor, D. T. Toledano, M. Bautista, S. King, D. Wang, and J. Colas, "Augmented set of features for confidence estimation in spoken term detection," in *Interspeech*, 2010.

[160] J. Tejedor, D. Torre, M. Bautista, and J. Colas, "Speech signal- and term-based feature contribution to hit/false alarm classification in a spoken term detection system," in *FALA*, 2010.

[161] N. Kanda, R. Takeda, and Y. Obuchi, "Using rhythmic features for japanese spoken term detection," in *SLT*, 2012.

[162] T. Ohno and T. Akiba, "Incorporating syllable duration into line-detection-based spoken term detection," in *SLT*, 2012.

[163] S. Goldwater, D. Jurafsky, and C. D. Maning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, pp. 181 – 200, 2009.

[164] C. Ma and C.-H. Lee, "A study on word detector design and knowledge-based pruning and rescoring," in *Interspeech*, 2007.

[165] J. Tejedor, A. Echeverria, and D. Wang, "An evolutionary confidence estimation for spoken term detection," in *CBMI*, 2011.

[166] O. Kurland, L. Lee, and C. Domshlak, "Better than the real thing?: iterative pseudo-query processing using cluster-based language models," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005.

[167] T. Tao and C. Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.

[168] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008.

[169] 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.

[170] ——, "Positional relevance model for pseudo-relevance feedback," in *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

[171] W.-H. Lin, R. Jin, and A. Hauptmann, "Web image retrieval re-ranking with relevance model," in *Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*, 2003.

[172] A. P. Natsev, M. R. Naphade, and J. TešiĆ, "Learning the semantics of multimedia queries and concepts from a small number of examples,"

in *Proceedings of the 13th annual ACM international conference on Multimedia*, ser. MULTIMEDIA '05, 2005, pp. 598–607.
[173] R. Yan, A. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003.
[174] S. Rudinac, M. Larson, and A. Hanjalic, "Exploiting visual reranking to improve pseudo-relevance feedback for spoken-content-based video retrieval," in *Image Analysis for Multimedia Interactive Services, 2009. WIAMIS '09. 10th Workshop on*, 2009.
[175] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proceedings of the 2nd international conference on Image and video retrieval*, ser. CIVR'03, 2003, pp. 238–247.
[176] R. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *Speech and Audio Processing, IEEE Transactions on*, 1996.
[177] H.-Y. Lee and L.-s. Lee, "Enhanced spoken term detection using support vector machines and weighted pseudo examples," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 1272–1284, 2013.
[178] I.-F. Chen and C.-H. Lee, "A hybrid HMM/DNN approach to keyword spotting of short words," in *Interspeech*, 13.
[179] A. Norouzian, A. Jansen, R. Rose, and S. Thomas, "Exploiting discriminative point process models for spoken term detection," in *Interspeech*, 2012.
[180] A. Norouzian, R. Rose, and A. Jansen, "Semi-supervised manifold learning approaches for spoken term verification," in *Interspeech*, 2013.
[181] H.-Y. Lee, P.-W. Chou, and L.-s. Lee, "Improved open-vocabulary spoken content retrieval with word and subword indexing using acoustic feature similarity," in *Special Issue on Information Extraction and Retrieval, Computer Speech and Language*, 2014, pp. 1045–1065.
[182] ——, "Open-vocabulary retrieval of spoken content with shorter/longer queries considering word/subword-based acoustic feature similarity," in *Interspeech*, 2012.
[183] 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.
[184] H.-Y. Lee, "Spoken content retrieval – relevance feedback, graphs and semantics," Ph.D. dissertation, National Taiwan University, 2012.
[185] A. Norouzian, R. Rose, S. H. Ghalehjegh, and A. Jansen, "Zero resource graph-based confidence estimation for open vocabulary spoken term detection," in *ICASSP*, 2013.
[186] J. Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, pp. 888–905, 2000.
[187] A. N. Langville and C. D. Meyer, "A survey of eigenvector methods for web information retrieval," *SIAM Rev.*, vol. 47, pp. 135–161, January 2005.
[188] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 17, pp. 107 – 117, 1998.
[189] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking through random walk over document-level context graph," in *Proceedings of the 15th international conference on Multimedia*, 2007, pp. 971–980.
[190] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *Proceedings of the 16th ACM international conference on Multimedia*, 2008, pp. 131–140.
[191] J. Otterbacher, G. Erkan, and D. R. Radev, "Biased lexrank: Passage retrieval using random walks with question-based priors," *Information Processing & Management*, vol. 45, no. 1, pp. 42 – 54, 2009.
[192] 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.
[193] *http://cool.mcgill.ca/*.
[194] F. Metze, N. Rajput, X. Anguera, M. Davel, G. Gravier, C. van Heerden, G. Mantena, A. Muscariello, K. Prahallad, I. Szoke, and J. Tejedor, "The spoken web search task at mediaeval 2011," in *ICASSP*, 2012.
[195] F. Metze, E. Barnard, M. Davel, C. V. Heerden, X. Anguera, G. Gravier, and N. Rajput., "The spoken web search task," in *MediaEval 2012 Workshop*, 2012.
[196] X. Anguera, F. Metze, A. Buzo, I. Szoke, and L. J. Rodriguez-Fuentes, "The spoken web search task," in *MediaEval 2013 Workshop*, 2013.
[197] X. Anguera, L.-J. R. Fuentes, I. Szke, A. Buzo, and F. Metze, "Query by example search on speech at mediaeval 2014," in *MediaEval 2014 Workshop*, 2014.
[198] *http://www.multimediaeval.org/*.
[199] F. Metze, X. Anguera, E. Barnard, M. Davel, and G. Gravier, "Language independent search in mediaeval's spoken web search task," *Computer Speech & Language*, vol. 28, no. 5, pp. 1066 – 1082, 2014.
[200] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, pp. 43–49, 1978.
[201] M. Mueller, *Information Retrieval for Music and Motion*. Springer-Verlag, 2007, ch. 4, pp. 69–84.
[202] A. Park and J. Glass, "Unsupervised pattern discovery in speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 186–197, Jan 2008.
[203] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on gaussian posteriorgrams," in *ICASSP*, 2010.
[204] C.-A. Chan, "Unsupervised spoken term detection with spoken queries," Ph.D. dissertation, National Taiwan University, 2012.
[205] X. Anguera and M. Ferrarons, "Memory efficient subsequence DTW for Query-by-Example spoken term detection," in *ICME*, 2013.
[206] I. Szoke, L. Burget, F. Grezl, and L. Ondel, "BUT SWS 2013 - massive parallel approach," in *MediaEval*, 2013.
[207] M. Calvo, M. Gimenez, L.-F. Hurtado, E. Sanchis, and J. A. Gomez, "ELiRF at MediaEval 2014: Query by example search on speech task (QUESST)," in *MediaEval*, 2014.
[208] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *ASRU*, 2009.
[209] Y. Zhang and J. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *ICASSP*, 2010.
[210] H. Wang and T. Lee, "CUHK system for QUESST task of MediaEval 2014," in *MediaEval*, 2014.
[211] ——, "CUHK system for the spoken web search task at MediaEval 2012," in *MediaEval*, 2012.
[212] J. Tejedor, M. Fapšo, I. Szöke, J. H. Černocký, and F. Grézl, "Comparison of methods for language-dependent and language-independent query-by-example spoken term detection," *ACM Trans. Inf. Syst.*, 2012.
[213] I. Szoke, L. Burget, F. Grezl, J. Cernocky, and L. Ondel, "Calibration and fusion of query-by-example systems – BUT SWS 2013," in *ICASSP*, 2014.
[214] S. Kesiraju, G. Mantena, and K. Prahallad, "IIIT-H system for MediaEval 2014 QUESST," in *MediaEval*, 2014.
[215] G. V. Mantena and K. Prahallad, "IIIT-H SWS 2013: Gaussian posteriorgrams of bottle-neck features for query-by-example spoken term detection," in *MediaEval*, 2013.
[216] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "An acoustic segment modeling approach to query-by-example spoken term detection," in *ICASSP*, 2012.
[217] C.-Y. Lee and J. Glass, "A nonparametric bayesian approach to acoustic model discovery," in *ACL*, 2012.
[218] Y. Zhang, R. Salakhutdinov, H. Chang, and J. Glass, "Resource configurable spoken query detection using deep Boltzmann machines," in *ICASSP*, 2012.
[219] A. Jansen and P. Niyogi, "Intrinsic spectral analysis," *Signal Processing, IEEE Transactions on*, 2013.
[220] P. Yang, C.-C. Leung, L. Xie, B. Ma, and H. Li, "Intrinsic spectral analysis based on temporal context features for query by example spoken term detection," in *Interspeech*, 2014.
[221] M. A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, "Rapid evaluation of speech representations for spoken term discovery," in *Interspeech*, 2011.
[222] M. R. Gajjar, R. Govindarajan, and T. V. Sreenivas, "Online unsupervised pattern discovery in speech using parallelization," in *Interspeech*, 2008.
[223] Y. Zhang and J. R. Glass, "Fast spoken query detection using lower-bound dynamic time warping on graphical processing units," in *ICASSP*, 2012.
[224] C.-A. Chan and L.-s. Lee, "Unsupervised spoken term detection with spoken queries using segment-based dynamic time warping," in *Interspeech*, 2010.
[225] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," in *ICASSP*, 2008.
[226] E. Keogh, "Exact indexing of dynamic time warping," in *Proc. of the 28th international conference on Very Large Data Bases*, 2002, pp. 406–417.
[227] Y. Zhang and J. Glass, "An inner-product lower-bound estimate for dynamic time warping," in *ICASSP*, 2011.
[228] ——, "A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping," in *Interspeech*, 2011.

[229] A. Jansen and B. V. Durme, "Efficient spoken term discovery using randomized algorithms," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[230] ——, "Indexing raw acoustic feature for scalable zero resource search," in *Interspeech*, 2012.

[231] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, 2002, pp. 380–388.

[232] X. Anguera, "Information retrieval-based dynamic time warping," in *Interspeech 2013*, 2013.

[233] A. Jansen and K. Church, "Towards unsupervised training of speaker independent acoustic models," in *Interspeech*, 2011.

[234] A. Jansen, K. Church, and H. Hermansky, "Towards spoken term discovery at scale with zero resources," in *Interspeech*, 2010.

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

[236] L. Wang, E. S. Chng, and H. Li, "An iterative approach to model merging for speech pattern discovery," in *APSIPA*, 2011.

[237] N. Vanhainen and G. Salvi, "Word discovery with beta process factor analysis," in *Interspeech*, 2012.

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

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

[240] C.-T. Chung, C.-A. Chan, and L.-s. Lee, "Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization," in *ICASSP*, 2013.

[241] A. Jansen, S. Thomas, and H. Hermansky, "Weak top-down constraints for unsupervised acoustic model training," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013.

[242] L. J. Rodrguez-Fuentes, A. Varona, M. Peagarikano, G. Bordel, and M. Dez, "GTTS systems for the SWS task at MediaEval 2013." in *MediaEval*, 2013.

[243] L. J. Rodriguez-Fuentes, A. Varona, M. Penagarikano, G. Bordel, and M. Diez, "GTTS-EHU systems for QUESST at MediaEval 2014," in *MediaEval*, 2014.

[244] H. Wang, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Using parallel tokenizers with DTW matrix combination for low-resource spoken term detection," in *ICASSP*, 2013.

[245] P. Yang, H. Xu, X. Xiao, L. Xie, C.-C. Leung, H. Chen, J. Yu, H. Lv, L. Wang, S. J. Leow, B. Ma, E. S. Chng, and H. Li, "The NNI query-by-example system for MediaEval 2014," in *MediaEval*, 2014.

[246] A. Buzo, H. Cucu, and C. Burileanu, "SpeeD @ MediaEval 2014: Spoken term detection with robust multilingual phone recognition," in *MediaEval*, 2014.

[247] J. Proena, A. Veiga, and F. Perdigao, "The SPL-IT query by example search on speech system for MediaEval 2014," in *MediaEval*, 2014.

[248] M. Huijbregts, M. McLaren, and D. van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," in *ICASSP*, 2011.

[249] C.-A. Chan and L.-s. Lee, "Model-based unsupervised spoken term detection with spoken queries," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 1330–1342, 2013.

[250] C.-A. Chan, C.-T. Chung, Y.-H. Kuo, and L.-s. Lee, "Toward unsupervised model-based spoken term detection with spoken queries without annotated data," in *ICASSP*, 2013.

[251] C.-A. Chan and L.-s. Lee, "Unsupervised hidden markov modeling of spoken queries for spoken term detection without speech recognition," in *Interspeech*, 2011.

[252] R. A. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[253] C.-T. Chung, C.-A. Chan, and L.-s. Lee, "Unsupervised spoken term detection with spoken queries by multi-level acoustic patterns with varying model granularity," in *ICASSP*, 2014.

[254] C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[255] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613–620, 1975.

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

[257] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *SIGIR*, 1998.

[258] C. Zhai and J. Lafferty, "A study of smoothing methods for language models applied to ad hoc information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001.

[259] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, pp. 391–407, 1990.

[260] T. Hofmann, "Probabilistic latent semantic analysis," in *In Proc. of Uncertainty in Artificial Intelligence, UAI99*, 1999.

[261] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[262] D. D. Lee and H. S. Seun, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[263] A. Singhal and F. Pereira, "Document expansion for speech retrieval," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[264] B. Billerbeck and J. Zobel, "Document expansion versus query expansion for ad-hoc retrieval," in *Proceedings of the Tenth Australasian Document Computing Symposium*, 2005.

[265] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.

[266] Q. Wang, J. Xu, H. Li, and N. Craswell, "Regularized latent semantic indexing," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

[267] X. Yi and J. Allan, "A comparative study of utilizing topic models for information retrieval," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, 2009.

[268] D. Metzler and W. B. Croft, "Latent concept expansion using Markov random fields," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[269] V. Lavrenko and W. B. Croft, "Relevance based language models," in *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001.

[270] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *CIKM*, 2001.

[271] B. Chen, P.-N. Chen, and K.-Y. Chen, "Query modeling for spoken document retrieval," in *ASRU*, 2011.

[272] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART retrieval system - experiments in automatic document processing*, 1971.

[273] 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.

[274] Y. chi Li and H. M. Meng, "Document expansion using a side collection for monolingual and cross-language spoken document retrieval," in *In ISCA Workshop on Multilingual Spoken Document Retrieval*, 2003, pp. 85–90.

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

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

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

[278] 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.

[279] 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.

[280] T. K. Chia, H. Li, and H. T. Ng, "A statistical language modeling approach to lattice-based spoken document retrieval," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[281] H.-Y. Lee and L.-s. Lee, "Improved semantic retrieval of spoken content by document/query expansion with random walk over acoustic similarity graphs," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, pp. 80–94, 2014.

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

[283] J. S. Olsson, "Vocabulary independent discriminative term frequency estimation," in *Interspeech*, 2008.

[284] 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.

[285] 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.

[286] J. Mamou and B. Ramabhadran, "Phonetic query expansion for spoken document retrieval," in *Interspeech*, 2008.

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

[288] P.-N. Chen, K.-Y. Chen, and B. Chen, "Leveraging relevance cues for improved spoken document retrieval," in *Interspeech*, 2011.

[289] B. Chen, H. min Wang, and L.-s. Lee, "Discriminating capabilities of syllable-based features and approaches of utilizing them for voice retrieval of speech information in mandarin chinese," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, pp. 303–314, 2002.

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

[291] H.-Y. Lee, Y.-C. Li, C.-T. Chung, and L.-s. Lee, "Enhancing query expansion for semantic retrieval of spoken content with automatically discovered acoustic patterns," in *ICASSP*, 2013.

[292] J. Glass, "Towards unsupervised speech processing," in *Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on*, July 2012, pp. 1–4.

[293] Y.-C. Li, H.-Y. Lee, C.-T. Chung, C.-A. Chan, and L.-s. Lee, "Towards unsupervised semantic retrieval of spoken content with query expansion based on automatically discovered acoustic patterns," in *ASRU*, 2013.

[294] D. W. Oard, J. White, J. Paik, R. Sankepally, and A. Jansen, "The FIRE 2013 question answering for the spoken web task," in *FIRE*, 2013.

[295] T. J. Hazen, F. Richardson, and A. Margolis, "Topic identification from audio recordings using word and phone recognition lattices," in *ASRU*, 2007.

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

[297] J. Marston, G. MacCarthy, B. Logan, P. Moreno, and J.-M. Van Thong, "News tuner: a simple interface for searching and browsing radio archives," in *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, 2004.

[298] M. Light and M. T. Maybury, "Personalized multimedia information access," *Commun. ACM*, vol. 45, pp. 54–59, 2002.

[299] S.-Y. Kong and L.-s. Lee, "Semantic analysis and organization of spoken documents based on parameters derived from latent topics," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 1875 –1889, 2011.

[300] 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, no. 2, pp. 632–645, 2012.

[301] K. Riedhammer, M. Gropp, and E. Noth, "The FAU video lecture browser system," in *SLT*, 2012.

[302] H.-Y. Lee, S.-R. Shiang, C.-F. Yeh, Y.-N. Chen, S.-Y. K. Yu Huang, and L.-s. Lee, "Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 5, pp. 881–896, May 2014.

[303] G. Tur and R. DeMori, *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*. John Wiley & Sons Inc, 2011, ch. 13.

[304] S. R. Maskey, "Automatic broadcast news speech summarization," Ph.D. dissertation, Columbia University, 2008.

[305] S.-H. Lin, B. Chen, and H.-M. Wang, "A comparative study of probabilistic ranking models for chinese spoken document summarization," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, pp. 3:1–3:23, 2009.

[306] S.-H. Lin, Y.-M. Chang, J.-W. Liu, and B. Chen, "Leveraging evaluation metric-related training criteria for speech summarization," in *ICASSP*, 2010.

[307] X. Cai and W. Li, "Ranking through clustering: An integrated approach to multi-document summarization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, pp. 1424–1433, 2013.

[308] S. Xie and Y. Liu, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, vol. 24, pp. 495 – 514, 2010.

[309] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *ASRU*, 2009.

[310] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A global optimization framework for meeting summarization," in *ICASSP*, 2009.

[311] F. Liu and Y. Liu, "Towards abstractive speech summarization: Exploring unsupervised and supervised approaches for spoken utterance compression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 7, pp. 1469–1480, July 2013.

[312] Y. Fujii, K. Yamamoto, N. Kitaoka, and S. Nakagawa, "Class lecture summarization taking into account consecutiveness of important sentences," in *Interspeech*, 2008.

[313] 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, no. 4, pp. 401–408, 2004.

[314] J. Zhang, R. Chan, and P. Fung, "Extractive speech summarization using shallow rhetorical structure modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, pp. 1147 –1157, 2010.

[315] D. Harwath and T. J. Hazen, "Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech," in *ICASSP*, 2012.

[316] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *ICASSP*, 2008.

[317] H.-Y. Lee, Y.-N. Chen, and L.-s. Lee, "Utterance-level latent topic transition modeling for spoken documents and its application in automatic summarization," in *ICASSP*, 2012.

[318] S.-Y. Kong and L.-s. Lee, "Improved spoken document summarization using probabilistic latent semantic analysis (PLSA)," in *ICASSP*, 2006.

[319] N. Garg, B. Favre, K. Reidhammer, and D. Hakkani-Tur, "Clusterrank: A graph based method for meeting summarization," in *Interspeech*, 2009.

[320] J. Zhang, H. Y. Chan, P. Fung, and L. Cao, "A comparative study on speech summarization of broadcast news and lecture speech," in *Interspeech*, 2007.

[321] H.-Y. Lee, Y.-Y. Chou, Y.-B. Wang, and L.-s. Lee, "Supervised spoken document summarization jointly considering utterance importance and redundancy by structured support vector machine," in *Interspeech*, 2012.

[322] ——, "Unsupervised domain adaptation for spoken document summarization with structured support vector machine," in *ICASSP*, 2013.

[323] S.-R. Shiang, H.-Y. Lee, and L.-s. Lee, "Supervised spoken document summarization based on structured support vector machine with utterance clusters as hidden variables," in *Interspeech*, 2013.

[324] S. Xie, B. Favre, D. Hakkani-Tur, and Y. Liu, "Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization," in *Interspeech*, 2009.

[325] S. Xie and Y. Liu, "Using confusion networks for speech summarization," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

[326] K. Zechner and A. Waibel, "Minimizing word error rate in textual summaries of spoken language," in *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 2000.

[327] H. Nanjo and T. Kawahara, "A new ASR evaluation measure and minimum bayes-risk decoding for open-domain speech understanding," in *ICASSP*, 2005.

[328] A. Inoue, T. Mikami, and Y. Yamashita, "Improvement of speech summarization using prosodic information," in *Proc. Speech Prosody*, 2004.

[329] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Interspeech*, 2005.

[330] S. Xie, D. Hakkani-Tur, B. Favre, and Y. Liu, "Integrating prosodic features in extractive meeting summarization," in *ASRU*, 2009.

[331] N. Chen, B. Ma, and H. Li, "Minimal-resource phonetic language models to summarize untranscribed speech," in *ICASSP*, 2013.

[332] X. Zhu, G. Penn, and F. Rudzicz, "Summarizing multiple spoken documents: finding evidence from untranscribed audio," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, 2009.

[333] S.-Y. Kong, C.-C. Wang, K.-C. Kuo, and L.-s. Lee, "Automatic title generation for Chinese spoken documents with a delicate scored Viterbi algorithm," in *SLT*, 2008.

[334] E. D'Avanzo, B. Magnini, and A. Vallin, "Keyphrase extraction for summarization purposes: The LAKE system at DUC-2004," in *Proceedings of the 2004 document understanding conference*, 2004.
[335] X. Jiang, Y. Hu, and H. Li, "A ranking approach to keyphrase extraction," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009.
[336] F. Liu, F. Liu, and Y. Liu, "Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion," in *SLT*, 2008.
[337] ——, "A supervised framework for keyword extraction from meeting transcripts," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, pp. 538–548, 2011.
[338] F. Liu, D. Pennell, F. Liu, and Y. Liu, "Unsupervised approaches for automatic keyword extraction using meeting transcripts," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009.
[339] T. J. Hazen, "Latent topic modeling for audio corpus summarization," in *Interspeech*, 2011.
[340] Y.-N. Chen, Y. Huang, S.-Y. Kong, and L.-s. Lee, "Automatic key term extraction from spoken course lectures using branching entropy and prosodic/semantic features," in *SLT*, 2010.
[341] Y.-N. Chen, Y. Huang, H.-Y. Lee, and L.-s. Lee, "Unsupervised two-stage keyword extraction from spoken documents by topic coherence and support vector machine," in *ICASSP*, 2012.
[342] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, and A. Saarela, "Self organization of a massive document collection," *Neural Networks, IEEE Transactions on*, vol. 11, pp. 574–585, 2000.
[343] T. Hofmann, "ProbMap – a probabilistic approach for mapping large document collections," *Intell. Data Anal.*, vol. 4, pp. 149–164, 2000.
[344] *http://www.google.com/insidesearch/features/search/knowledge.html*.
[345] F. Kubala, S. Colbath, D. Liu, and J. Makhoul, "Rough'n'Ready: a meeting recorder and browser," *ACM Comput. Surv.*, vol. 31, no. 2es, Jun. 1999. [Online]. Available: http://doi.acm.org/10.1145/323216.323354
[346] A. G. Hauptmann, "Lessons for the future from a decade of informedia video analysis research," in *Proceedings of the 4th International Conference on Image and Video Retrieval*, 2005.
[347] T.-H. Li, M.-H. Lee, B. Chen, and L.-s. Lee, "Hierarchical topic organization and visual presentation of spoken documents using probabilistic latent semantic analysis (PLSA) for efficient retrieval/browsing applications," in *Interspeech*, 2005.
[348] C. Carpineto, S. Osiński, G. Romano, and D. Weiss, "A survey of web clustering engines," *ACM Comput. Surv.*, vol. 41, pp. 17:1–17:38, 2009.
[349] Y.-C. Pan, J.-Y. Chen, Y.-S. Lee, Y.-S. Fu, and L.-s. Lee, "Efficient interactive retrieval of spoken documents with key terms ranked by reinforcement learning," in *Interspeech*, 2006.
[350] Y.-C. Pan and L.-s. Lee, "Simulation analysis for interactive retrieval of spoken documents with key terms ranked by reinforcement learning," in *SLT*, 2006.
[351] I. Ruthven, *Interactive information retrieval*. John Wiley & Sons, Inc., 2008, vol. 42, no. 1.
[352] T. Misu and T. Kawahara, "Bayes risk-based dialogue management for document retrieval system with speech interface," *Speech Communication*, vol. 52, no. 1, pp. 61 – 71, 2010.
[353] T.-H. Wen, H.-Y. Lee, and L.-S. Lee, "Interactive spoken content retrieval with different types of actions optimized by a markov decision process," in *Interspeech*, 2012.
[354] J. D. Williams and S. Young, "Partially observable markov decision processes for spoken dialog systems," *Computer Speech & Language*, vol. 21, pp. 393 – 422, 2007.
[355] J. Liu, P. Pasupat, S. Cyphers, and J. Glass, "ASGARD: a portable architecture for multilingual dialogue systems," in *ICASSP*, 2013.
[356] E. Levin, R. Pieraccini, and W. Eckert, "A stochastic model of human-machine interaction for learning dialog strategies," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, pp. 11–23, 2000.
[357] C. Szepesvári and R. Munos, "Finite time bounds for sampling based fitted value iteration," in *ICML*, 2005.
[358] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-based user simulation for bootstrapping a pomdp dialogue system," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, 2007.
[359] T.-H. Wen, H.-Y. Lee, P.-H. Su, and L.-S. Lee, "Interactive spoken content retrieval by extended query model and continuous state space Markov decision process," in *ICASSP*, 2013.
[360] S. Young, "Talking to machines (statistically speaking)," in *ICSLP*, 2002.
[361] ——, "Still talking to machines (cognitively speaking)," in *Interspeech*, 2010.
[362] O. Lemon, "Learning what to say and how to say it: Joint optimisation of spoken dialogue management and natural language generation," *Computer Speech & Language*, vol. 25, pp. 210 – 221, 2011.
[363] H.-Y. Lee, Y. Zhang, E. Chuangsuwanich, and J. Glass, "Graph-based re-ranking using acoustic feature similarity between search results for spoken term detection on low-resource languages," in *Interspeech*, 2014.

Lin-shan Lee (F93) 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 on the Board of Governors (1995), the Vice President for International Affairs (19961997) and the Awards Committee chair (19981999) of the IEEE Communications Society. He was a member of the Board of International Speech Communication Association (ISCA 20022009), a Distinguished Lecture (20072008) and a member of the Overview Paper Editorial Board (2009-2010) 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, and IEEE ComSoc/KICS Exemplary Global Service Award from IEEE Communication Society in 2014.

James Glass is a Senior Research Scientist at the Massachusetts Institute of Technology where he leads the Spoken Language Systems Group in the Computer Science and Artificial Intelligence Laboratory. He is also a member of the Harvard-MIT Health Sciences and Technology Faculty. Since obtaining his Ph.D. at MIT in Electrical Engineering and Computer Science, his research has focused on automatic speech recognition, unsupervised speech processing, and spoken language understanding. He is an Associate Editor for the IEEE Transactions on Audio, Speech, and Language Processing, and a member of the Editorial Board for Computer, Speech, and Language. He is also an IEEE Fellow, and a Fellow of the International Speech Communication Association.

Hung-yi Lee, received the M.S. and Ph.D. degrees 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. From September 2013 to July 2014, he was a visiting scientist at the Spoken Language Systems Group of MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). He is currently an assistant professor of the Department of Electrical Engineering of National Taiwan University, with a joint appointment at the Department of Computer Science & Information Engineering of the university. His research focuses on spoken language understanding, speech recognition and machine learning.

Chun-an Chan received his Ph.D. degree in communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2012. He spent another year as a postdoctoral researcher in the Speech Processing Lab in 2013. He is currently a software engineer in Google, working on text-to-speech synthesis technology.