Integrating Recognition and Retrieval With Relevance Feedback for Spoken Term Detection

Hung-yi Lee, Chia-ping Chen, and Lin-shan Lee, *Fellow, IEEE*

***Abstract*—Recognition and retrieval are typically viewed as two cascaded independent modules for spoken term detection (STD). Retrieval techniques are assumed to be applied on top of automatic speech recognition (ASR) output, with performance depending on ASR accuracy. We propose a framework that integrates recognition and retrieval and consider them jointly in order to yield better STD performance. This can be achieved either by adjusting the acoustic model parameters (model-based) or by considering detected examples (example-based) using relevance information provided by the user (user relevance feedback) or inferred by the system (pseudo-relevance feedback), either for a given query (short-term context) or by taking into account many previous queries (long-term context). Such relevance feedback approaches have long been used in text information retrieval, but are rarely considered and cannot be directly applied to the retrieval of spoken content. The proposed relevance feedback approaches are specific to spoken content retrieval and are hence very different from those developed for text retrieval, which are applied only to text symbols. We present not only these relevance feedback scenarios and approaches for STD, but also propose a framework to integrate them all together. Preliminary experiments showed significant improvements in each case.**

***Index Terms*—Relevance feedback, spoken term detection.**

I. Introduction

In the Internet era, digital content over the Internet covers all the information and activities of human life. The most attractive form of network content is multimedia, which commonly includes speech. The subjects, topics, and core concepts of such multimedia content can very often be identified based on the speech information within the content. Hence, in the future, speech information retrieval will be very important in helping users retrieve and browse efficiently the huge qualities of multimedia content [1]. In general, there are two stages in conventional speech information retrieval approaches [2]. In the first stage, the audio content is recognized and transformed into transcriptions or lattices by a recognition engine based on a set of acoustic models and language models. In the second stage, after the user enters a query, the retrieval engine searches through the recognition output and returns a list of relevant spoken segments to the user. The returned segments are usually ranked by the relevance scores derived from the recognition output. Here we focus on spoken term detection (STD), in which the query is a term in text form and a spoken segment is considered relevant if it includes the query term. These discussions may be generalized to other tasks in speech information retrieval as well.

Substantial effort has been made in speech information retrieval, and many successful techniques have been developed. Lattice-based approaches taking into account multiple recognition hypotheses [3], [4] have been used to mitigate the relatively low accuracy in 1-best transcriptions. Lattices are usually converted into sausage-like structures to make the indexing task easier and reduce memory requirements. Examples of such sausage-like lattice-based structures include position specific posterior lattices (PSPL) [5], [6] and confusion networks (CN) [6], [7]. Out-of-vocabulary (OOV) queries is another important issue because typically many queries contain OOV terms [8]. The most common approach for handling the OOV problem is to represent both the queries and the spoken segments by properly chosen subword units and then match them at the subword unit level [9]–[15]. Word-based and subword-based indexing can be further integrated to yield improved performance [9], [16], [17]. Various successful applications of spoken content retrieval have been demonstrated including those for broadcast news [18], course lectures [19], [20], podcasts [21], and YouTube videos [22].

In the past, recognition and retrieval have been treated as two cascaded independent modules; thus, the assumption was that they should be individually considered and optimized. Because most spoken term detection techniques were developed to be applied on top of ASR output—either 1-bests or lattices—a common assumption is that retrieval performance depends heavily on ASR accuracy. Clearly, there are limitations when the retrieval process is simply to be applied on top of the ASR output. For poor recognition accuracies, even if the correct word hypotheses can be included in the lattice, the low posterior probabilities may make it difficult to detect the spoken terms covered by many incorrect noisy terms. Therefore, spoken term detection performance is inevitably dominated by ASR performance. However, in many practical applications, it is difficult to obtain acoustic and language models robust enough for the huge quantities of target spoken segments generated by many speakers under different conditions for different

Manuscript received September 11, 2011; revised January 31, 2012; accepted April 11, 2012. Date of publication April 25, 2012; date of current version June 11, 2012. This work was supported in part by the National Science Council, China, under Contract NSC 97-2221-E-002-134-MY3 and in part by the Ministry of Education under Contract 99R80303. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gokhan Tur.

H.-Y. Lee is with the Department of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: tlkagkb93901106@gmail.com; tlkagkb93901106@yahoo.com.tw).

C.-P. Chen was with the Department of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan. She is now with MediaTek, Inc., Hsinchu 30078, Taiwan (e-mail: coward7652@yahoo.com.tw).

L.-S. Lee is with the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: lslee@gate.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2196514

applications with different domains. In such cases even very robust retrieval approaches are not able to compensate for the recognition errors.

Some researchers have jointly taken into account the recognition and retrieval processes in an attempt to improve retrieval performance. One good example is considering the recognition error pattern with a confusion matrix during retrieval [23], [24]. This involves inferring the correct words actually appearing in the spoken segments from the erroneous ASR transcriptions. Some have also observed that although word accuracy is an excellent metric for recognition performance, it is not directly related to retrieval performance [25]–[27]. For example, words frequently used as query terms should be correctly recognized, while recognition errors for function words have almost no impact on retrieval performance. As a result, word significance has been taken into account during decoding [25], [26], and a minimum classification error (MCE [28]) discriminative training method was used that also took into account word significance [27]. In another approach, when an OOV query term is entered, the OOV term is dynamically inserted into the possible positions in the lattice to handle the OOV query [29].

On the other hand, it has been known for long in text information retrieval that an effective way of improving the retrieval performance is to involve the users into the search process by relevance feedback. During the search process, the user provides information about relevant segments as positive examples and irrelevant segments as negative examples with respect to the query entered. Thus, the system learns from these examples to yield improved performance. The relevance information may come from the user (**user relevance feedback**), but it can also be inferred by the system automatically without involving the user (**pseudo-relevance feedback**). Additionally, there are two scenarios for user relevance feedback: *short-term* and *long-term* context. These will be detailed below. Relevance feedback is a mature technique for text information retrieval [30], [31], and it has been applied on numerous popular text retrieval models such as the vector space model [30], the probability model [32], and the language model [33]. It has been used extensively in different retrieval domains such as image [34]–[37] and video retrieval [38]–[40]; however, it has not yet been fully leveraged for speech information retrieval.

We propose integrating the recognition and retrieval modules as a whole, and further enhancing STD performance with various relevance feedback-based approaches. The basic concept of this proposed approach is shown in Fig. 1. As in standard STD approaches, the baseline system is a cascade of the recognition and retrieval systems shown in the lower left part of the figure. Each spoken segment in the archive (Fig. 1, middle) is first transcribed into a lattice by a recognition engine based on a set of acoustic models. When a query is entered by a user, the retrieval system searches over the lattices, and returns a ranked list of matching spoken segments (lower right) to the user. The user then gives feedback to the system, for example by selecting items 1 and 3 as relevant but item 2 as irrelevant, as shown in the upper right corner; again, this kind of information could also be generated automatically without actually involving the user. We propose integrating the recognition and retrieval modules using a feedback loop, with the goal

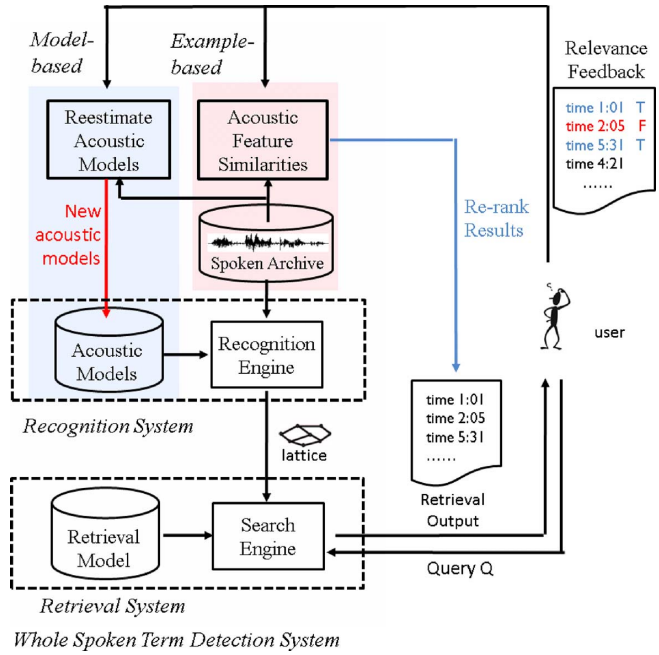

Fig. 1. Framework of the proposed approaches.

of jointly improving the two modules. First, the parameters of the acoustic models used for recognition can be adjusted according to the feedback relevancy information. The posterior probabilities of the hypotheses on the lattice for spoken segments in the first-pass returned list are then updated based on the new set of acoustic models, and the relevance scores are modified accordingly. Thus, the recognition and retrieval modules are jointly optimized. This approach is a **model-based** approach [41], [42], shown in the upper-left corner of the figure. Retrieval performance can be improved also by taking into account the similarity in acoustic features between the spoken segments in the first-pass returned list and those with relevance information. Those hypotheses with acoustic features similar to or different from the positive examples may thus be detected from the returned list. This is an **example-based** approach [43], [44], shown in the upper middle. Such model- and example-based approaches can be used with either user relevance feedback or pseudo-relevance feedback. We further introduce a new framework in which all the different scenarios and approaches of relevance feedback for STD can be integrated together [45].

The relevance feedback approaches proposed here, whether adjusting acoustic model parameters (model-based) or considering detected examples (example-based), are specifically designed for spoken content retrieval and are thus quite different from those used for text information retrieval. Although the concept of relevance feedback has been thoroughly studied for text information retrieval, such approaches by their nature take into account only text information. For speech information retrieval, it is true that these text-based methods can be directly applied to audio archive transcriptions. However, the transcriptions may include many recognition errors which imply information loss, that is, when transcribing speech signals into text, much information is lost and not recoverable. Therefore, a better approach is not directly applying these relevance feedback approaches on the transcriptions, but on the speech signal level

(example-based) or recognition model level (model-based) with a hope to better use those information carried by the signals, which may be lost during recognition. This is why in this paper we propose novel example-based and model-based relevance feedback approaches for spoken term detection, which are quite different from those used on text information retrieval, and test them on typical feedback scenarios to explore their effectiveness and limitations. The approaches proposed here may render STD performance less dependent on the acoustic models, which are often mismatched to the spoken archive. Clearly the spoken segments available from many different web sites over the Internet exhibit widely variant acoustic and linguistic conditions; thus it can be almost impossible to get matched data for acoustic and language model adaptation. The approaches proposed here are thus an important step towards more robust STD technologies.

In Section II, we summarize different scenarios for relevance feedback, and in Section III we introduce the relevance score functions used in STD tasks. Example- and model-based relevance feedback approaches are respectively presented in Sections IV and V, and in Section VI we summarize all of the scenarios and approaches and propose a framework under which to integrate them. In Sections VII–IX, we report the experimental results. We conclude in Section X.

II. Scenarios of Relevance Feedback

When a query Q is entered, the retrieval system in Fig. 1 ranks the returned spoken segments X based on the values of a relevance score function $S(Q,X)$ evaluated for X with respect to Q. Relevance information obtained from the feedback loop modify the original score $S(Q,X)$ to yield a better score $S'(Q,X)$. As mentioned above, there are in general two different scenarios of relevance feedback: user relevance feedback and pseudo-relevance feedback. These are further discussed in this section.

A. User Relevance Feedback

Although in this scenario the relevance information comes from the user, implicit feedback [46]–[49] has been widely used in real systems because most users are reluctant to give relevance feedback explicitly. Implicit feedback means the system analyzes the user's behavior online to get the feedback information; the user does not know he is in a feedback procedure. One example is click-through data [47]. We assume that the transcription is displayed beside each spoken segment on the returned list given by the retrieval system, and that the user is able to judge if the segment is what he wants based on the automatic transcriptions. It may be reasonable to assume that the user only clicks on the segments considered relevant. Thus, if a user clicks on the third spoken segment on the returned list without clicking on the first two, it is reasonable for the system to assume that the third segment is relevant and that the first two are irrelevant.

Such user relevance feedback includes short- and long-term context scenarios [49], [50]. For short-term context user relevance feedback, the retrieval system obtains relevance information only for the single query a user just entered, and the relevance feedback process attempts only to improve the retrieval

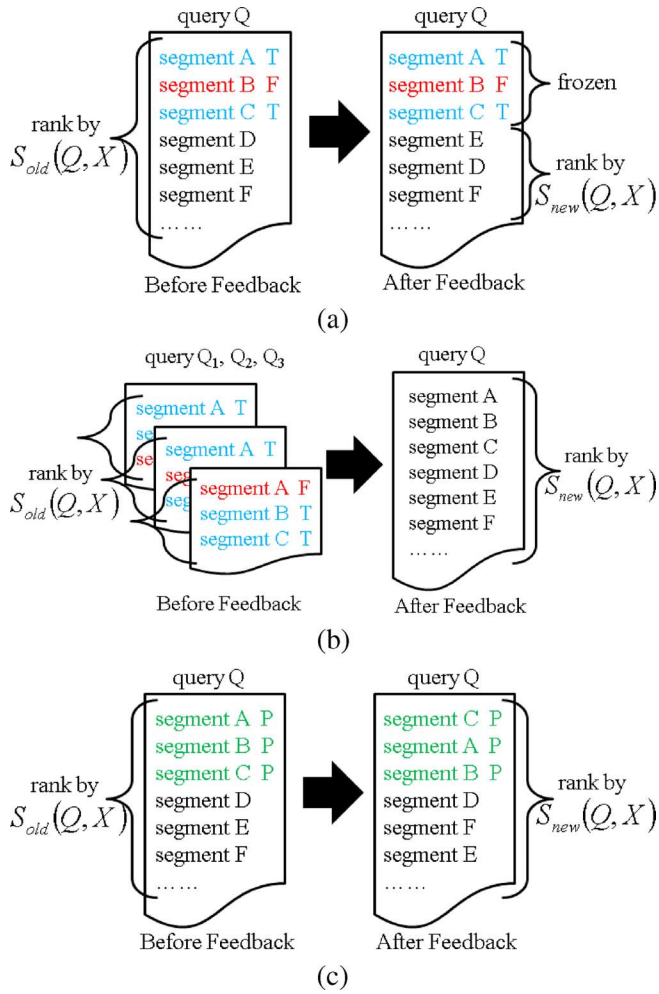

Fig. 2. Different relevance feedback scenarios. The original score $S(Q,X)$ before relevance feedback is changed to $S'(Q,X)$ after relevance feedback. Spoken segments with T, F and P are respectively the user-labeled relevant and irrelevant segments, and those assumed relevant by the system. (a) Short-term context user relevance feedback. (b) Long-term context user relevance feedback. (c) Pseudo-relevance feedback.

performance for exactly the current query. For long-term context user relevance feedback, the historical record of relevance information for many different queries is used to improve the retrieval performance over all other queries. These two scenarios are further discussed next.

1) Short-Term Context User Relevance Feedback: Fig. 2(a) shows short-term context user relevance feedback. The user browses the retrieved list on the left side ranked by the original score $S(Q,X)$. If the user gives the relevance information of the top N spoken segments on the list to the system, those labeled segments are used to obtain the new score $S'(Q,X)$, which is used to re-rank the spoken segments below the top N. Note that the order of these top N labeled spoken segments should be frozen [51]. In practice, the returned results are usually divided into pages. When the user clicks through the first page, he actually gives relevance information implicitly to the system. When he starts to browse the second page, the system has already changed the ranking order of the spoken segments after the first page based on the new score which includes the relevance information from the first page. In this case, the top N spoken segments with user relevance information are the

spoken segments in the first page, and because the user has already seen them, re-ranking their order is meaningless, and thus they should be frozen.

2) Long-Term Context User Relevance Feedback: Fig. 2(b) shows long-term context user relevance feedback. Historical relevance information for many queries [training queries such as Q_1, Q_2, and Q_3 in Fig. 2(b)] entered by one or more users is collectively used to train the new score, $S'(Q,X)$, which is used to rank the spoken segments corresponding to the new query Q'.

B. Pseudo-Relevance Feedback

Pseudo-relevance feedback [38], [39], [52]–[61] is widely used to obtain relevance information without actually involving users. The basic idea of pseudo-relevance feedback is to assume that a small number of top-ranked objects in the initial returned results are relevant (or "pseudo-relevant"), and the user does not actually participate in the feedback process in any way. As shown in Fig. 2(c), the system simply assumes the top M spoken segments in the first-pass returned list ranked by $S(Q,X)$ are relevant without any user input [$M = 3$ in Fig. 2(c)]; these pseudo-relevance segments are then used as positive examples to obtain $S'(Q,X)$ as in the short-term context. All of the returned spoken segments are then re-ranked based on this new score. Note that no spoken segments' orders should be frozen here because no spoken segment has been labeled by the user. In fact, what is presented to the user is the re-ranked list of spoken segments after pseudo-relevance feedback.

III. Lattice-based Relevance Score Function

For STD, after the user enters a text query Q, the retrieval engine searches through the recognition output for the spoken archive and returns a list of relevant spoken segments. Here, all spoken segments X in the archive are ranked by their degree of relevance with respect to the query Q, represented by relevance score function $S(Q,X)$. In this section, we briefly introduce this score used in the baseline STD system before any relevance feedback, which is derived from the lattices. This score is widely used in STD [62]–[67].

The audio signal is first divided into spoken segments, after which each segment X in the spoken archive is transcribed to a lattice $W(X)$. When the query Q (either a word or a phrase) is entered, all the spoken segments X in the spoken archive are ranked based on

$$S(Q,X|\theta) = \frac{\sum_{u\in W(X)} P_\theta(X|u)P(u)N(u,Q)}{\sum_{u\in W(X)} P_\theta(X|u)P(u)} \quad (1)$$

where u is an allowed word sequence in the lattice $W(X)$, $P_\theta(X|u)$ is the likelihood for observation sequence X given the word sequence u based on the acoustic model set θ, $P(u)$ is the prior probability of u from the language model, and $N(u,Q)$ is the occurrence count of query Q in u. Since the denominator in (1) is the sum of the likelihoods of all word sequences u in the lattice, and the numerator of (1) is the same but weighted by the occurrence count of query Q, (1) can be interpreted as the expected occurrence count of query Q in lattice $W(X)$ based on the set of acoustic models θ. We include the set of acoustic model parameters θ in the function because in Section V a new set of acoustic model parameters θ^* will be obtained via relevance feedback, which will lead to a new relevance score function in (1).

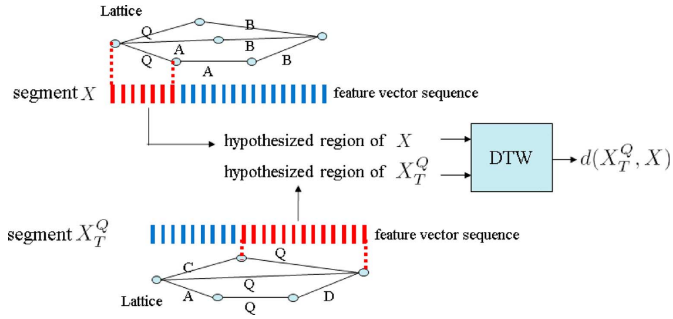

Fig. 3. The hypothesized region is defined as the corresponding time span of a word arc in the lattice whose word hypothesis is exactly the query term Q with the highest posterior probability in the lattice. The distance of a spoken segment X and a positive example X_T^Q is based on the dynamic time warping (DTW) distance between the MFCC sequences for their hypothesized regions.

IV. Example-Based Relevance Feedback

In this approach, we assume that a given term is always pronounced the same way and thus exhibits similar acoustic feature sequences. Thus, if positive examples for the desired query are available, it is possible to judge the correctness of the query hypotheses in the retrieved spoken segments based on the similarity of the acoustic feature sequences of these query hypotheses to those of the given positive examples. These positive examples can be obtained from either user relevance feedback or pseudo-relevance feedback. In this approach, we first define the "hypothesized region" for a spoken segment X with respect to a query Q to be the corresponding time span of a word arc in the lattice for the spoken segment whose word hypothesis is exactly the query term Q with the highest posterior probability, as shown in Fig. 3. Thus, if a spoken segment has a "hypothesized region" very similar to those of known positive examples, it is more likely to be relevant, so its relevance score should be increased. This approach can be used in either the short-term context scenario (Section II-A1) or the pseudo-relevance feedback scenario (Section II-B). It cannot be applied in the long-term context scenario (Section II-A2) because the examples of a specific query term cannot be generalized to other queries.

A. Short-Term Context User Relevance Feedback

Given the positive example set $\mathcal{S}_T^Q$ of spoken segments with respect to the query Q annotated by the user and collected by the system in the short-term context, the similarity between the whole set and a spoken segment X in the first-pass returned list is computed based on their hypothesized regions. We first define the distance between such a spoken segment X and the whole positive example set $\mathcal{S}_T^Q$ as

$$D(\mathcal{S}_T^Q, X) = \sum_{X_T^Q \in \mathcal{S}_T^Q} d(X_T^Q, X)^b \quad (2)$$

where X_T^Q is a positive example segment in the set $\mathcal{S}_T^Q$, and $d(X_T^Q, X)$ is the dynamic time warping (DTW) distance be-

tween the two feature vector sequences corresponding to the two hypothesized regions of X_T^Q and X as shown in Fig. 3, the summation in (2) is over all positive examples, and b is a parameter scaling the values of DTW distances. When performing DTW, Euclidean distance or any other distances can be used to compute the distance between the two feature vectors. We can then transform this distance $D(\mathcal{S}_T^Q, X)$ in (2) into a similarity measure between $\mathcal{S}_T^Q$ and X

$$SIM(\mathcal{S}_T^Q, X) = 1 - \frac{D(\mathcal{S}_T^Q, X)}{M_Q} \tag{3}$$

where M_Q is the maximum value of $D(\mathcal{S}_T^Q, X)$ for all spoken segments X in the first-pass returned list for query Q. Although we here report only experiments using MFCC features for the DTW distances, other speech frame representations could be used, such as Gaussian posteriorgrams, which provide less speaker-dependent DTW distance measures [68], [69] and could thus be useful if the target spoken segments are produced by many different speakers. With (3), we linearly normalize the score into a range between zero and one representing the similarity. This normalization is used for simplicity; other normalization functions are possible. Finally, we integrate the similarity obtained in (3) with the original $S(Q, X|\theta)$ to obtain $S_1'(Q, X|\theta)$ and then re-rank the spoken segments in the first-pass returned list accordingly:

$$S_1'(Q, X|\theta) = S(Q, X|\theta)(SIM(\mathcal{S}_T^Q, X))^a \tag{4}$$

where a is a weighting parameter.

B. Pseudo-Relevance Feedback

For pseudo-relevance feedback, everything is exactly the same as described above for short-term context except that the positive examples are now replaced by those obtained via pseudo-relevance feedback, referred to as pseudo positive examples. The system simply assumes that the top M spoken segments in the returned list ranked by $S(Q, X|\theta)$ are relevant. Therefore, the positive example set $\mathcal{S}_T^Q$ in (4) is replaced by the pseudo positive example set $\mathcal{S}_P^Q$, and the spoken segments in the fist-pass returned list are now re-ranked according to

$$S_2'(Q, X|\theta) = S(Q, X|\theta)(SIM(\mathcal{S}_P^Q, X))^{a'} \tag{5}$$

where a' is another weighting parameter.

V. Model-Based Relevance Feedback

In model-based relevance feedback approach, a new set of acoustic models is reestimated based on the relevance information. Thus, because the relevance score function in (1) depends on the acoustic model parameters, it is changed accordingly, which in turn yields new ranking results.

Estimating acoustic model parameters based on predefined criterion is a well-studied problem in speech recognition. Applied to STD with relevance feedback, however, the problem is different from the conventional acoustic model training approaches for speech recognition in at least two ways:

1) The system input includes only whether a spoken segment is relevant to a query or not; it does not include the transcription of any utterance [21], [70].
2) The goal is to improve retrieval performance rather than recognition accuracy.

In this section, we propose a set of objective functions that take into account the retrieval process, as well as discriminative training algorithms that optimize these objective functions. Below, we show the model-based method can be used in both short-term (Section II-A1) and long-term (Section II-A2) contexts. Pseudo-relevance feedback (Section II-B) also seems possible, although adjusting model parameters based simply on assumed pseudo-relevance information may be risky.

A. Short-Term Context User Relevance Feedback

1) Objective Function: Given positive and negative (or relevant and irrelevant) examples for a certain query Q from the user relevance feedback, the system estimates a new set of acoustic model parameters θ^* by maximizing an objective function $F(\theta)$:

$$\theta^* = \arg\max_{\theta} F(\theta). \tag{6}$$

With the new set of acoustic models, the likelihood $P_\theta(X|u)$ in (1) is replaced by $P_{\theta^*}(X|u)$, so the original relevance score function in (1) for each segment is modified accordingly to $S(Q, X|\theta^*)$, based on which all the segments in the first-pass returned list are re-ranked. The above procedure is conducted online. It is not very time consuming because only a limited amount of data is used for model training. The new acoustic models are stored only in memory and are discarded after the retrieval session. Several objective functions $F(\theta)$ in (6) are proposed below.

The first objective function $F_1^Q(\theta)$ to be maximized in (6) is the sum of the relevance scores of all positive examples

$$F_1^Q(\theta) = \sum_{X_T^Q} S(Q, X_T^Q|\theta) \tag{7}$$

where X_T^Q is a positive example with respect to the query Q. The second objective function $F_2^Q(\theta)$ is then the sum of the distances between all positive and negative example pairs

$$F_2^Q(\theta) = \sum_{X_T^Q, X_F^Q} [S(Q, X_T^Q|\theta) - S(Q, X_F^Q|\theta)] \tag{8}$$

where X_F^Q is a negative example with respect to the query Q.

Since mean average precision (MAP) [71] widely used in many STD tasks is used here as the basic measure to evaluate retrieval performance in the experiments, maximizing the distances between all pairs of positive/negative examples as in (8) does not necessarily yield improved retrieval performance. MAP quantifies the goodness of the ranked retrieval results, and as such favors retrieval results with relevant documents ranked higher than irrelevant documents. Thus, the relative levels of all positive examples with respect to all negative examples are more important than their individual absolute relevance score differences. To be specific, if a positive example already has a

higher relevance score than all negative examples, any increase in the relevance score of this positive example cannot further benefit retrieval performance. Therefore, the second objective function $F_2^Q(\theta)$ is not effective enough. The acoustic model training process can be enhanced if we can estimate a set of acoustic models that directly maximizes the MAP of the training examples. Although directly optimizing MAP may be difficult, it has been found that maximizing an accuracy count $A(\theta)$ is equivalent to maximizing a lower MAP bound [47], [72]:

$$A(\theta) = \sum_{X_T^Q, X_F^Q} \delta(X_T^Q, X_F^Q) \tag{9}$$

where

$$\delta(X_T^Q, X_F^Q) = \begin{cases} 1, & S(X_T^Q, Q|\theta) > S(X_F^Q, Q|\theta) \\ 0, & \text{otherwise} \end{cases}. \tag{10}$$

$A(\theta)$ hence represents the number of positive/negative example pairs in which the relevance of the positive example is greater than that of the negative example. However, since $A(\theta)$ in (9) is not differentiable, it is not easily optimized. Therefore, we approximate $\delta(X_T^Q, X_F^Q)$ in (10) with

$$sigmoid(X_T^Q, X_F^Q) = \frac{1}{1 + e^{c[S(Q, X_T^Q|\theta) - S(Q, X_F^Q|\theta)]}} \tag{11}$$

and define the third objective function to be optimized as

$$F_3^Q(\theta) = \sum_{X_T^Q, X_F^Q} sigmoid(X_T^Q, X_F^Q). \tag{12}$$

In (11), as $S(X_T^Q, Q|\theta)$ is larger than $S(X_F^Q, Q|\theta)$, $sigmoid(X_T^Q, X_F^Q)$ tends to 1; otherwise, $sigmoid(X_T^Q, X_F^Q)$ tends to 0. c is a constant that controls the slope of the sigmoid function.

When utilizing (12) as the objective function, the estimated acoustic model may overfit to the training examples. For instance, the acoustic models may rank all positive examples higher than all negative examples, but it is possible that some positive examples may be scored lower than some unlabeled segments, since the unlabeled data is not considered at all in (12). This may not be good because some of these unlabeled segments may be irrelevant. Hence, we wish to estimate a set of acoustic models which keeps the positive examples ranked at the top of the first-pass returned list, including those unlabeled, to prevent such overfitting. This can be achieved by replacing the objective function $F_3^Q(\theta)$ with

$$F_4^Q(\theta) = F_3^Q(\theta) + \rho \sum_{X_T^Q, X_{un}^Q} sigmoid(X_T^Q, X_{un}^Q) \tag{13}$$

where X_{un}^Q is an unlabeled segment within the returned list and ρ is a weighting parameter. $sigmoid(X_T^Q, X_{un}^Q)$ tends to 1 if X_T^Q has a higher relevance score than X_{un}^Q. Equation (13) can be viewed as a smoothing approach that ensures the unlabeled segments are given lower scores than the positive examples.

2) Optimization: All the objective functions presented in Section V-A1 can be optimized using the weak-sense auxiliary function similar to that in minimum phone error (MPE) discriminative training [73]. MPE maximizes the expected phone accuracy as

$$F_{MPE}(\theta) = \sum_{r=1}^{R} \frac{\sum_{u \in W(X_r)} P_\theta(X_r|u) P(u) A(u)}{\sum_{u \in W(X_r)} P_\theta(X_r|u) P(u)} \tag{14}$$

where X_r is the rth training utterance, R is the total number of training utterances, $A(u)$ is the phone accuracy evaluated for the corresponding phone sequence of the word sequence u, while everything else has the same definition as in (1). Taking $F_2^Q(\theta)$ in (8) as an example, here we first show that objective functions $F_1^Q(\theta)$ and $F_2^Q(\theta)$ mentioned in Section V-A1 can be manipulated to have the same form as (14) except for a word sequence u with a different definition of $A(u)$.

Recall that the relevance score function in (1) is written as

$$S(Q, X|\theta) = \frac{\sum_{u \in W(X)} P_\theta(X|u) P(u) N(u, Q)}{\sum_{u \in W(X)} P_\theta(X|u) P(u)} \tag{15}$$

where $N(u, Q)$ is the occurrence count of the word hypothesis Q in the word sequence u. Hence, substituting (15) into (8) yields

$$\begin{aligned} F_2^Q(\theta) = &\sum_{X_T^Q} \frac{\sum_{u \in W(X_T^Q)} P_\theta(X_T^Q|u) P(u) |X_F^Q| N(u, Q)}{\sum_{u \in W(X_T^Q)} P_\theta(X_T^Q|u) P(u)} \\ &+ \sum_{X_F^Q} \frac{\sum_{u \in W(X_F^Q)} P_\theta(X_F^Q|u) P(u) |X_T^Q| N'(u, Q)}{\sum_{u \in W(X_F^Q)} P_\theta(X_F^Q|u) P(u)} \end{aligned} \tag{16}$$

where $W(X_T^Q)$, $W(X_F^Q)$ are the sets of all possible word sequences in the lattices for the examples X_T^Q and X_F^Q, respectively, $|X_T^Q|$ and $|X_F^Q|$ are the total number of positive and negative examples included in the evaluation in (8), and $N'(u, Q)$ is defined as $-N(u, Q)$. Therefore, we can optimize (16) in exactly the same way as for MPE by simply replacing $A(u)$ in (14) by $|X_F^Q| N(u, Q)$ or $|X_T^Q| N'(u, Q)$ as in (16). Note that just like in MPE, in the model estimation process, the acoustic models are updated iteratively starting from an initial acoustic model set.

The optimization of $F_3^Q(\theta)$ in (12) is more complicated. In the MPE model estimation process, at the ith iteration, given the acoustic model set θ_{i-1} obtained in the last iteration, a new acoustic model set θ_i maximizing a weak-sense auxiliary function of (14) is estimated. The auxiliary function used in MPE training is

$$\begin{aligned} &H_{MPE}(\theta_i, \theta_{i-1}) \\ &\quad = \sum_{r=1}^{R} \sum_{a \in A(X_r)} \left[\frac{\partial F_{MPE}(\theta_{i-1})}{\partial \log P_{\theta_{i-1}}(X_r|a)} \right] \log P_{\theta_i}(X_r|a) \end{aligned} \tag{17}$$

where $A(X_r)$ represents all the arcs in the lattice of utterance X_r, and $\partial F_{MPE}(\theta_{i-1}) / \partial \log P_{\theta_{i-1}}(X_r|a)$ is a constant with

TABLE I
SUMMARY OF DIFFERENT RELEVANCE FEEDBACK SCENARIOS AND APPROACHES

Scenarios		Approaches	
		Example-based (Sec IV)	Model-based (Sec V)
User relevance feedback	short-term (Sec II-A1)	case (α)	case (γ)
	long-term (Sec II-A2)	N/A	case (ϵ)
Pseudo-relevance feedback	short-term (Sec II-B)	case (β)	-

respect to the acoustic models θ_i to be estimated. $F_3^Q(\theta)$ in (12) can be optimized in a similar way. At the ith training iteration, we find for $F_3^Q(\theta)$ in (12) the auxiliary function

$$H_{F_3}(\theta_i, \theta_{i-1}) = \sum_{X_T^Q} \sum_{a \in A(X_T^Q)} \left[\frac{\partial F_3^Q(\theta_{i-1})}{\partial \log P_{\theta_{i-1}}(X_T^Q|a)} \right] \log P_{\theta_i}(X_T^Q|a) + \sum_{X_F^Q} \sum_{a \in A(X_F^Q)} \left[\frac{\partial F_3^Q(\theta_{i-1})}{\partial \log P_{\theta_{i-1}}(X_F^Q|a)} \right] \log P_{\theta_i}(X_F^Q|a) \tag{18}$$

where $A(X_T^Q)$ and $A(X_F^Q)$ represent all the arcs in the lattices of utterances X_T^Q and X_F^Q. Then the new acoustic model θ_i maximizing (18) can be estimated in exactly the same way that (17) is maximized in MPE discriminative training. The optimization of $F_4^Q(\theta)$ in (13) is then trivial.

B. Long-Term Context User Relevance Feedback

One of the strengths of the model-based approach is that it can be used in a long-term context for which the example-based approach is not applicable. In long-term context user relevance feedback, the system collects a set of training queries $\mathcal{Q}_{train} = \{Q_1, Q_2, Q_3, \ldots\}$ and their positive and negative examples. The retrieval system can therefore estimate a new set of acoustic model parameters θ_{lt}^* by maximizing

$$F_5^{lt}(\theta) = \sum_{Q \in \mathcal{Q}_{train}} F^Q(\theta) \tag{19}$$

which is the summation over the objective functions of all the queries in the training query set Q_{train}. F^Q in (19) can be $F_1^Q(\theta)$ in (7), $F_2^Q(\theta)$ in (8), $F_3^Q(\theta)$ in (12), or $F_4^Q(\theta)$ in (13). The new models θ_{lt}^* are then used to rescore all the lattices in the spoken archive, and then the lattices with new scores are stored and indexed for further use. This approach can yield overall improvements to system performance, even for queries that were not included in the training query set.

VI. SUMMARY OF DIFFERENT RELEVANCE FEEDBACK SCENARIOS AND APPROACHES AND A FRAMEWORK OF FURTHER INTEGRATION

Here, we summarize all the above relevance feedback scenarios and approaches and propose a framework to properly integrate them together.

A. Summary of Scenarios and Approaches

The above relevance feedback scenarios and approaches are first summarized in Table I. The table rows correspond to the scenarios discussed in Section II, including user relevance feedback and pseudo-relevance feedback, where the former may be based on either short- or long-term context. The columns correspond to the different approaches (example-based in Section IV or model-based in Section V). In the example-based approach, the new relevance score function is $S_1'(Q, X|\theta)$ for short-term context user relevance feedback in (4) or $S_2'(Q, X|\theta)$ for pseudo-relevance feedback in (5), in which the original score $S(Q, X|\theta)$ is multiplied by similarity measure $SIM(\mathcal{S}_T^Q, X)^a$ or $SIM(\mathcal{S}_P^Q, X)^{a'}$. For the model-based approach, the score is $S(Q, X|\theta^*)$ as in (1) except that a new set of acoustic model parameters θ^* is used, and θ^* is optimized with an objective function $F(\theta)$ as in (6). There are six cells in Table I, four of which are labeled as cases (α), (β), (γ), and (ϵ). We focus on these four cases here and in the experimental results below. In principle, model-based approaches can be performed with pseudo-relevance feedback. However, since the pseudo-positive examples are already the segments with the highest relevance scores, estimating a new set of acoustic models based on some criteria maximizing their relevance scores may have little effect for the purpose here because the original acoustic model parameters already maximize their relevance scores. Due to lack of space, the experimental results below do not include model-based approaches for pseudo-relevance feedback. Note that long-term context user relevance feedback cannot be performed with example-based approaches.

B. Framework Integrating Different Scenarios and Approaches

In Fig. 4, is shown a framework integrating the above scenarios and approaches in an STD system, in which the conventional STD system is the block at the middle left. The rationale by which we integrate different scenarios and approaches is as follows. If historical relevance information is available, long-term context user relevance feedback should be applied offline first to improve the system performance overall before a new query is entered. For long-term context, only the model-based approach applies: case (ϵ) in the upper left. Given the user query in the bottom right, pseudo-relevance feedback without real user interaction is applied before the returned list is shown to the user. Here, we only apply the example-based approach for pseudo-relevance feedback: case (β) in the middle. Therefore, the retrieval results the user obtains after he enters his query are actually the results that have already been improved by pseudo-relevance feedback. Then, the user has an opportunity to provide to the system relevance information for some segments. Thus, short-term context user relevance feedback can be applied at this moment to further improve the retrieval results, in which both example-based [case (α)] and model-based

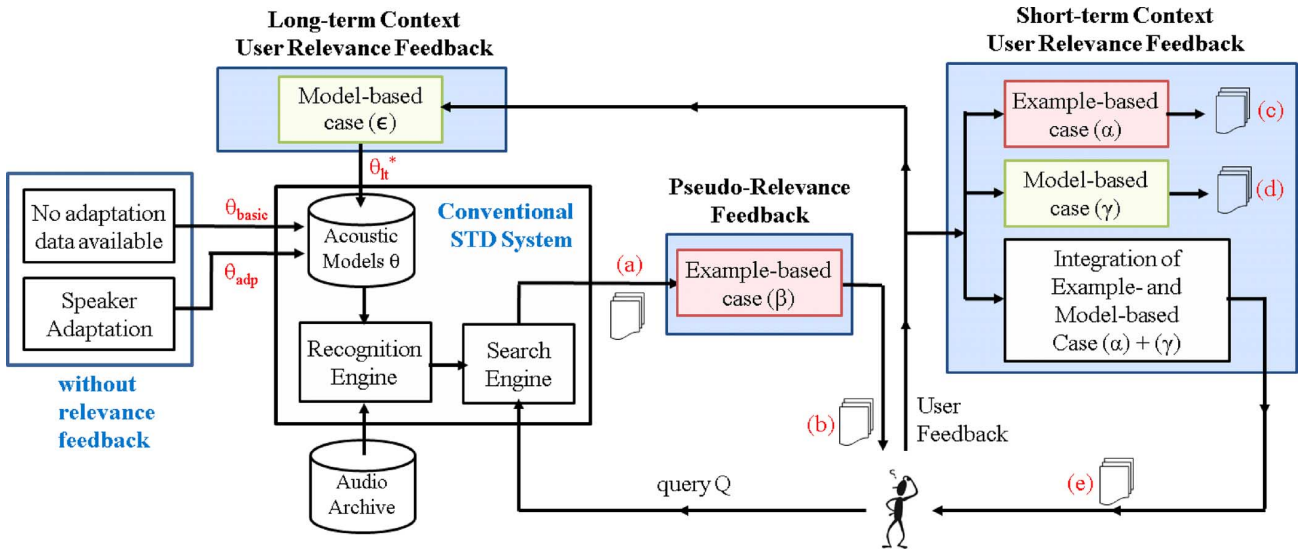

Fig. 4. Framework integrating different STD relevance feedback scenarios and approaches. Lists (a) to (e) correspond to the retrieval results of different integration configuration with relevance score functions listed in Table II.

TABLE II
RELEVANCE SCORE FUNCTIONS USED FOR THE FIVE DIFFERENT LISTS OF RESULTS (a) TO (e) UNDER DIFFERENT INTEGRATION CONFIGURATION AS SHOWN IN FIG. 4. THE ACOUSTIC MODEL PARAMETER SET θ HERE CAN BE θ_{basic} (UNADAPTED), θ_{adp} (ADAPTED), OR θ^*_{lt} (ESTIMATED WITH LONG-TERM RELEVANCE INFORMATION)

list	relevance score function	cases applied
(a)	$S(Q,X\|\theta)$	None
(b)	Equation (20)	case (β)
(c)	Equation (21)	case (β) + (α)
(d)	Equation (22)	case (β) + (γ)
(e)	Equation (23)	case (β) + (α) + (γ)

[case (γ)] approaches can be used, and they can be further integrated [case $(\alpha]$ plus case $[\gamma)]$, as in the right part of Fig. 4.

This rationale leads to a discussion of the relevance score function used in each case. Given query Q, we first generate the baseline results (list (a), middle of Fig. 4), ranked using the original score $S(Q,X|\theta)$ in (1). There can be three different cases for the acoustic model parameter set θ used here in the relevance score function. Because adaptation data matched to the spoken archive are usually not available, the lattices can only be generated by a set of task-independent acoustic models θ_{basic} which may be mismatched to the target spoken archive. In the rare case that adaptation data are available, a set of adapted acoustic model parameters θ_{adp} may be used. If the relevance information for previously entered queries are available, model-based long-term context user relevance feedback (case (ϵ), upper left corner) can be performed to yield the set θ^*_{lt}. Hence, the set θ here can be θ_{basic}, θ_{adp}, or θ^*_{lt}, as shown in the upper-left corner.

Before the retrieval result is shown to the user, the system first conducts example-based pseudo-relevance feedback [case (β)] as in the middle of Fig. 4 on list (a) to produce the list (b) at the lower middle. Here the baseline [list (a)] is re-ranked because the score is changed from $S(Q,X|\theta)$ to $S'_2(Q,X|\theta)$ in (5):

$$S'_2(Q,X|\theta) = S(Q,X|\theta)SIM(\mathcal{S}_P^Q,X)^{a'}. \tag{20}$$

This result after pseudo-relevance feedback [list (b)] is displayed to the user. Given user relevance information for spoken segments in list (b), the system conducts short-term context user relevance feedback by the example-based [case (α)] and/or model-based [case (γ)] methods in the upper right.

For example-based approaches, the segments in list (b) not fedback by the user are re-ranked as in (4) by the new score

$$\begin{aligned} S''_1(Q,X|\theta) &= S'_2(Q,X|\theta)SIM(\mathcal{S}_T^Q,X)^{a} \\ &= [S(Q,X|\theta)SIM(\mathcal{S}_P^Q,X)^{a'}]SIM(\mathcal{S}_T^Q,X)^{a} \end{aligned} \tag{21}$$

or score $S'_2(Q,X|\theta)$ for list (b) is multiplied by $SIM(\mathcal{S}_T^Q,X)^a$ [case (α)] to yield the new retrieval result [list (c)].

For model-based approaches, on the other hand, a new set of acoustic model parameters θ^* is estimated by maximizing one of the objective functions in Section V. In the model re-estimation process here the score $S(Q,X|\theta)$ to be used in optimizing $F(\theta)$ in (6) is replaced by $S(Q,X|\theta)SIM(\mathcal{S}_P^Q,X)^{a'}$, or $S'_2(Q,X|\theta)$ in (20), to take into account the influence of pseudo-relevance feedback. In this way, the new acoustic models θ^* focus on separating the relevance scores of the positive/negative example pairs which cannot be discriminated by example-based pseudo-relevance feedback. The segments in list (b) not fedback by the user are re-ranked by the new score

$$S'_2(Q,X|\theta^*) = S(Q,X|\theta^*)SIM(\mathcal{S}_P^Q,X)^{a'} \tag{22}$$

in which the new parameter θ^* is used, while everything else is the same as $S'_2(Q,X|\theta)$ in (20). This yields a new retrieval result [list (d)].

Finally, because both example-based [case (α)] and model-based [case (γ)] approaches can be applied in short-term context user relevance feedback, we integrate them by ranking those segments in list (b) not fedback by the user with the score

$$\begin{aligned} S''_1(Q,X|\theta^*) &= S'_2(Q,X|\theta^*)SIM(\mathcal{S}_T^Q,X)^{a} \\ &= [S(Q,X|\theta^*)SIM(\mathcal{S}_P^Q,X)^{a'}]SIM(\mathcal{S}_T^Q,X)^{a} \end{aligned} \tag{23}$$

which is exactly the same as (21) except that θ is replaced by θ^*. θ^* here is obtained by maximizing one of the objective functions in Section V with $S(Q, X|\theta)$ replaced by $S_1''(Q, X|\theta)$ in (21) to include the effect of both pseudo-relevance feedback and example-based user relevance feedback in the model re-estimation process. This is the final result [list (e)] that the system shows to the user after short-term context user relevance feedback. Equation (23) takes into account the new acoustic models θ^* obtained from model-based method as well as the similarities to positive examples. Although it is the result of integrating the model-based and example-based approaches [list (e)] that is shown to the user, the results of individually applying example-based [list (c)] and model-based methods [list (d)] for the short-term context user relevance feedback will also be reported in the experiments below. The relevance score functions used in the lists (a) to (e) are summarized in Table II. Note that in lists (c), (d), and (e), the cases (α) and/or (γ) are actually applied on top of case (β).

Long-term context user relevance feedback [case (ϵ)] in the upper left corner can be applied as well. That is, if the system has been online for a period of time, the historical user relevance feedback data can be collected to estimate a new set of parameters θ_{lt}^* with which the lattices for the spoken archive are rescored for further use. In this case, θ_{lt}^* replaces θ, and all the processes mentioned above can be repeated for new queries and relevance information given by new users.

VII. Experimental Setup

In this paper, we tested the proposed approaches on two different spoken archives. The first spoken archive we used was a set of recorded lectures for a university course (**Lecture**), and the second one was a broadcast news corpus (**News**). Mean average precision (MAP) was used as the retrieval performance evaluation measure. The pair-wise t-test with a significance level of 0.05 was used to gauge the significance of performance improvements. The experimental setups of the two testing spoken archives are described in the following two subsections.

A. Lecture

We used 33 hours of recorded lectures for a course offered in National Taiwan University as the first target spoken archive; it is quite noisy and spontaneous. The spoken archive was produced by a single instructor primarily in Mandarin Chinese but embedded with some English words. A Chinese lexicon with 10.7 K words and a phone set of 35 Mandarin phonemes (NTU-98 [74]) were used. Because of the lack of corpora matched to the topic (technical content of the course) and the style (spontaneous monologue) for the retrieved spoken archive here, the Chinese trigram language model was trained from the Mandarin Giga-word corpus released by Linguistic Data Consortium. Each spoken segment in the corpus was transcribed into a lattice with beamwidth of 50. Eighty Chinese queries were manually selected as testing queries, each consisting of a single word, and another twenty Chinese queries were used as a development set.

The unadapted acoustic models $\theta_{basic-l}$ (l indicates lecture here) used in the experiments for **Lecture** was trained using the maximum-likelihood criterion with 4602 state-tied triphones for Mandarin spanning 37 monophones using a corpus of clean read speech in Mandarin, which included 24.6 hours of data produced by 100 males and 100 females. 39-dimensional MFCCs were used as the feature. There were 5 states per triphone, and 24 mixtures per state. Both the acoustic and language models were highly mismatched, yielding a relatively poor character accuracy of 50.26%. For speaker adaptation, 500 utterances from the same target archive but not in the test set were used for adaptation. We tested two different speaker adapted models. Global MLLR was first applied to obtain the adapted acoustic models θ_{adp1} with character accuracy 62.55%. MLLR with 256 classes and maximum *a posteriori* estimation were then applied in addition to obtain a better set of adapted acoustic models θ_{adp2} with a character accuracy of 72.93%. Since the acoustic models above were based on Mandarin phonemes only, and the lexicon and language model used for recognizing **Lecture** only included Chinese words, the English words embedded in the speech were transcribed into Chinese word sequences with similar pronunciation, which made the retrieval task more challenging.

B. News

We used a broadcast news corpus in Mandarin Chinese as another spoken archive to test the proposed approaches. The news stories were recorded from TV stations in Taipei from 2001 to 2003, with a total length of 198 hours [74]. 160 Chinese queries were manually selected as testing queries, each consisting of a single word, and another 10 Chinese queries were used as a development set.

For the recognition of **News**, we used a 60 K-word lexicon, a tri-gram language model trained on 39 M words of Yahoo news, and a set of acoustic models $\theta_{basic-n}$ (n indicates news here) with 64 Gaussian mixtures per state and three states per model trained on a corpus of 24.5 hours of broadcast news different from the archive tested here. 147 right context-dependent initial models plus context-independent final models were used as the acoustic models for simplicity, and 39-dimensional MFCCs with cepstral mean and variance normalization (CMVN) applied were used as the features. Each spoken segment in the corpus was transcribed into a lattice with a beamwidth of 100. Since 48% and 31% of the speech in the corpus was produced by the reporters and respondents respectively including relatively high background noise, and only 147 acoustic models were used, the character accuracy for the archive was only 54.43%.

VIII. Experimental Results for Individual Relevance Feedback Cases

Here, we first present the experimental results of the two testing spoken archives, **Lecture** and **News**, for the four individual relevance feedback cases covering primary scenarios and approaches, or the cases (α), (β), (γ), and (ϵ) as summarized in Table I.

For **Lecture**, the unadapted acoustic models $\theta_{basic-l}$ and the speaker adapted acoustic models θ_{adp1} and θ_{adp2} were used to generate the lattices. The baseline retrieval system with models $\theta_{basic-l}$ without relevance feedback yielded an MAP score of 0.4819, and the MAP scores for θ_{adp1} and θ_{adp2} without relevance feedback were 0.6112 and 0.7307. On average, there were

respectively 120.5, 304.7, and 210.7 segments retrieved in the first-pass retrieval process for each query with models $\theta_{basic-l}$, θ_{adp1}, and θ_{adp2}. For **News**, the acoustic models $\theta_{basic-n}$ were used to generate the lattices, and the baseline retrieval system yielded an MAP score of 0.6302, and retrieved 443.3 segments for each query on average in the first-pass retrieval process. We observed that the accuracy of **News** was lower than the accuracy of **Lecture** with model θ_{adp1} (54.43% versus 62.55%), but **News** yielded a higher MAP (0.6302 versus 0.6112). This is probably because the **News** includes the very well produced, prepared speech by anchors and the very noisy, spontaneous speech by respondents and the interviews. The speech of the anchors was transcribed quite correctly, while the low accuracy of **News** primarily came from those of the respondents. The anchors' speech included more query terms which were reasonably well detected, therefore resulting in higher MAP even given the lower recognition accuracy of **News**.

A. Example-Based Relevance Feedback Approaches

Here we report the results of the example-based approach (Section IV) in the scenarios of short-term context user relevance feedback (case (α) in Table I) and pseudo-relevance feedback [case (β)]. Euclidean distance was used when computing the distance between MFCC vectors for DTW in (2) for all the experiments. The parameter b in (2) was set to 2.0 in all the experiments below.

1) Short-Term Context User Relevance Feedback (Case (α) in Table I): Here we consider example-based user relevance feedback in the short-term context. We compared the MAP for the testing queries before and after relevance feedback re-ranking. For each query, we assumed that the user provided the correct relevance information[1] for the top N (N =5,10,15,20) segments in the first-pass returned list ranked by $S(Q,X|\theta)$. θ could be either $\theta_{basic-l}$, θ_{adp1}, or θ_{adp2} for **Lecture**, or $\theta_{basic-n}$ for **News**. The segments below the top N were then re-ranked by modifying the original relevance score function $S(Q,X|\theta)$ by multiplying it with the similarity measure $SIM(\mathcal{S}_T^Q, X)^a$ as in (4), while the ranking of the top N segments in the list were frozen because they had been browsed by the user. The parameter a in (4) was decided by the development set.

Table III lists the results for N =5, 10, 15, and 20. The two sections **Lecture** and **News** correspond to the experimental results of the two testing spoken archives. The superscript labels $^{(0)}$ indicate significantly better than the baseline. It is interesting to note that larger N or more feedback segments sometimes offered less improvements in MAP. This is because MAP values are often dominated by the top several items selected, or the improvements in MAP scores were limited by the top N rank-frozen segments. In other words, when there were more labeled segments provided by the user, in general better relevance score function could be obtained, but the space left for improvements in MAP was also reduced, so the improvements achieved were not necessarily reflected in the MAP values. However, in all cases, the improvements in Table III were significant, except $N = 20$ with θ_{adp2} for **Lecture** and $N = 5$ for **News**.

[1]A segment is a positive or negative example, or the query term is in the spoken segment or not.

TABLE III
EXPERIMENTAL MAP RESULTS FOR EXAMPLE-BASED SHORT-TERM CONTEXT USER RELEVANCE FEEDBACK WHEN THE CORRECT RELEVANCE INFORMATION OF THE TOP N (N =5,10,15,20) SEGMENTS IN THE FIRST-PASS RETURNED LIST WAS GIVEN. THE TWO SECTIONS **LECTURE** AND **NEWS** CORRESPOND TO THE RESULTS OF TWO TESTING SPOKEN ARCHIVES. FOR **LECTURE**, THE THREE COLUMNS $\theta_{basic-l}$, θ_{adp1}, θ_{adp2} CORRESPOND TO THREE DIFFERENT ACOUSTIC MODELS USED FOR GENERATING THE LATTICES; FOR **NEWS**, THE LATTICES WERE GENERATED BY THE ACOUSTIC MODELS $\theta_{basic-n}$. THE SUPERSCRIPT LABELS $^{(0)}$ INDICATE SIGNIFICANTLY BETTER THAN THE BASELINE

Spoken Archive	**Lecture**			**News**
Acoustic Models	$\theta_{basic-l}$	θ_{adp1}	θ_{adp2}	$\theta_{basic-n}$
baseline	0.4819	0.6112	0.7307	0.6302
N=5	$0.5189^{(0)}$	$0.6471^{(0)}$	$0.7403^{(0)}$	0.6336
N=10	$0.5117^{(0)}$	$0.6446^{(0)}$	$0.7418^{(0)}$	$0.6377^{(0)}$
N=15	$0.5021^{(0)}$	$0.6386^{(0)}$	$0.7382^{(0)}$	$0.6377^{(0)}$
N=20	$0.4960^{(0)}$	$0.6331^{(0)}$	0.7369	$0.6361^{(0)}$

TABLE IV
EXPERIMENTAL MAP RESULTS OF EXAMPLE-BASED PSEUDO-RELEVANCE FEEDBACK WITH DIFFERENT NUMBER OF PSEUDO POSITIVE EXAMPLES M AND DIFFERENT ACOUSTIC MODELS USED FOR GENERATING THE LATTICES FOR THE TWO SPOKEN ARCHIVES. THE SUPERSCRIPT LABELS $^{(0)}$ INDICATE SIGNIFICANTLY BETTER THAN THE BASELINE

Spoken Archive	**Lecture**			**News**
Acoustic Models	$\theta_{basic-l}$	θ_{adp1}	θ_{adp2}	$\theta_{basic-n}$
baseline	0.4819	0.6112	0.7307	0.6302
M=1	$0.4984^{(0)}$	0.6112	0.7174	0.6163
M=3	$0.5062^{(0)}$	0.6168	0.7323	0.6318
M=5	$0.5166^{(0)}$	$0.6329^{(0)}$	0.7394	0.6302
M=7	$0.5188^{(0)}$	$0.6448^{(0)}$	$0.7420^{(0)}$	$0.6381^{(0)}$
M=9	$0.5208^{(0)}$	$0.6477^{(0)}$	$0.7423^{(0)}$	$0.6369^{(0)}$
M=11	$0.5192^{(0)}$	$0.6450^{(0)}$	0.7356	$0.6368^{(0)}$
M=13	$0.5158^{(0)}$	$0.6435^{(0)}$	0.7277	$0.6370^{(0)}$
M=15	$0.5132^{(0)}$	$0.6421^{(0)}$	0.7260	$0.6365^{(0)}$

2) Pseudo-Relevance Feedback (Case (β) in Table I): For example-based pseudo-relevance feedback, the pseudo positive segment set $\mathcal{S}_P^Q$ was simply the top M segments in the first-pass returned list with the highest $S(X,Q|\theta)$, and (5) was used to re-rank the segments with the parameter a' decided by the development set. M varied from 1 to 15. The results for different choices of M and different testing spoken archives with different acoustic models used for generating the lattices are shown in Table IV. We found that in most cases MAP first increased to a peak and then degraded as M was raised. The maximum improvements for **Lecture** for all different acoustic model sets were achieved when M was 9, while this number was 7 for **News**. This is reasonable because more relevant examples, or larger M, are helpful and the disturbance caused by incorrectly selected irrelevant segments can be diluted. However, when M was too large, more irrelevant segments are inevitably included, naturally degrading the MAP. M was thus set to 9 for **Lecture** and 7 for **News** in the following experiments. The maximum absolute improvements in MAP achieved with pseudo-relevance feedback on **Lecture** are 3.89%, 2.88%, and 1.16% for θ_{basic}, θ_{adp1}, and θ_{adp2}, and 0.79% for **News**.

B. Model-Based Relevance Feedback Approaches

Here we tested model-based approaches (Section V) in the scenarios of short- and long-term context user relevance

TABLE V
EXPERIMENTAL MAP RESULTS FOR **LECTURE** FOR MODEL-BASED SHORT-TERM CONTEXT USER RELEVANCE FEEDBACK WITH OBJECTIVE FUNCTIONS $F_1^Q(\theta)$, $F_2^Q(\theta)$, $F_3^Q(\theta)$ AND $F_4^Q(\theta)$ FOR N = 5,10,15,20. ACOUSTIC MODEL RE-ESTIMATION CAN BE STARTED WITH ACOUSTIC MODELS $\theta_{basic-l}$, θ_{adp1}, OR θ_{adp2}, AND THE MAP OF THE BASELINE WITHOUT RELEVANCE FEEDBACK ARE 0.4819, 0.6189 AND 0.7307 FOR LATTICES GENERATED BY $\theta_{basic-l}$, θ_{adp1}, AND θ_{adp2} RESPECTIVELY. THE SUPERSCRIPT LABELS $^{(0)}$, $^{(1)}$, $^{(2)}$ AND $^{(3)}$ RESPECTIVELY INDICATE SIGNIFICANTLY BETTER THAN THE BASELINE, $F_1^Q(\theta)$, $F_2^Q(\theta)$, AND $F_3^Q(\theta)$

Initial Acoustic Models	Number of Feedback segments	baseline	Objective Functions			
			$F_1^Q(\theta)$	$F_2^Q(\theta)$	$F_3^Q(\theta)$	$F_4^Q(\theta)$
$\theta_{basic-l}$	**N=5**	0.4819	0.4826	$0.5008^{(0)(1)}$	$0.5086^{(0)(1)(2)}$	$0.5106^{(0)(1)(2)}$
	N=10		0.4789	$0.5058^{(0)(1)}$	$0.5128^{(0)(1)(2)}$	$0.5140^{(0)(1)(2)}$
	N=15		0.4810	$0.5005^{(0)(1)}$	$0.5038^{(0)(1)}$	$0.5044^{(0)(1)(2)}$
	N=20		0.4813	$0.4998^{(0)(1)}$	$0.4990^{(0)(1)}$	$0.4998^{(0)(1)}$
θ_{adp1}	**N=5**	0.6189	0.6198	$0.6326^{(0)(1)}$	$0.6326^{(0)(1)}$	$0.6416^{(0)(1)(2)(3)}$
	N=10		0.6260	$0.6387^{(0)(1)}$	$0.6426^{(0)(1)(2)}$	$0.6485^{(0)(1)(2)(3)}$
	N=15		$0.6286^{(0)}$	$0.6287^{(0)}$	$0.6438^{(0)(1)(2)}$	$0.6427^{(0)(1)(2)}$
	N=20		$0.6293^{(0)}$	0.6244	$0.6387^{(0)(1)(2)}$	$0.6399^{(0)(1)(2)}$
θ_{adp2}	**N=5**	0.7307	0.7327	0.7366	$0.7443^{(0)(1)(2)}$	$0.7504^{(0)(1)(2)(3)}$
	N=10		0.7353	$0.7419^{(0)}$	$0.7431^{(0)(1)}$	$0.7492^{(0)(1)(2)(3)}$
	N=15		0.7351	0.7360	$0.7424^{(0)(1)(2)}$	$0.7461^{(0)(1)(2)}$
	N=20		0.7382	0.7372	$0.7421^{(0)(1)}$	$0.7416^{(0)(1)(2)}$

feedback [cases (γ) and (ϵ)]. For **Lecture**, the acoustic model re-estimation can be started with the unadapted acoustic models ($\theta_{basic-l}$) or the adapted models (θ_{adp1} or θ_{adp2}) used in generating the initial lattices; for **News**, the acoustic model re-estimation was started with the acoustic models $\theta_{basic-n}$. Again we assume the correct relevance information for the top N (N =5,10,15,20) segments were available. The user relevance feedback was used to re-estimate the acoustic model parameters including means, covariances, transition probabilities, and mixture weights.

1) Short-Term Context User Relevance Feedback (Case (γ) in Table I): Correct relevance information of the top N (N =5,10,15,20) segments was used in the model-based approach to obtain a new set of acoustic model parameters θ^* as in (6). The segments below the top N were then re-ranked based on the new score $S(Q, X|\theta^*)$, while the ranking of the top N segments were frozen. We compared the MAP scores of the returned list before and after re-ranking. All the smoothing parameters in the model training algorithm and the parameter ρ for $F_4^Q(\theta)$ in (13) were decided by the development set, and c in (11) was set to 1.0.

The experimental results of **Lecture** are shown in Table V for different objective functions ($F_1^Q(\theta)$, $F_2^Q(\theta)$, $F_3^Q(\theta)$, $F_4^Q(\theta)$ in (7), (8), (12), (13)) described in Section V-A with different N (N =5,10,15,20) and different initial models $\theta_{basic-l}$, θ_{adp1}, and θ_{adp2} used to generate the lattices. The new model parameter set θ^* was obtained with three training iterations. The superscripts labels on the MAP values, $^{(0)}$, $^{(1)}$, $^{(2)}$, and $^{(3)}$, respectively, indicate the MAP value is significantly better than the baseline, $F_1^Q(\theta)$, $F_2^Q(\theta)$, and $F_3^Q(\theta)$. As explained in Section VIII-A1, although more user labeled data (more training data) may lead to better acoustic models for the purpose here, the space left for improvements in MAP is reduced. Therefore, increasing the number of feedback segments N did not guarantee more improvements in MAP.

Much can be learned from Table V. First, it can be found that $F_2^Q(\theta)$ in (8) with the consideration of negative examples was always better than $F_1^Q(\theta)$ in (7) except when N = 20. Moreover, in all cases $F_2^Q(\theta)$ outperformed the baseline. $F_3^Q(\theta)$ always outperformed the baseline, $F_1^Q(\theta)$, and $F_2^Q(\theta)$ in all cases, except for N = 20 for $\theta_{basic-l}$. $F_4^Q(\theta)$ taking into account unlabeled data always outperformed $F_3^Q(\theta)$ in every case, except for N = 15 for θ_{adp1} and N = 20 for θ_{adp2}. $F_4^Q(\theta)$ did not outperform $F_3^Q(\theta)$ in those cases because $F_4^Q(\theta)$ was designed to handle the problem of overfitting, and therefore was of little benefit when N was large. These results in Table V verified that the considerations mentioned in Section V-A regarding $F_3^Q(\theta)$ and $F_4^Q(\theta)$ are all correct and contribute to the improvements. $F_4^Q(\theta)$ was found to be the best objective function, and with $F_4^Q(\theta)$ only five examples (N = 5) were needed to yield very significant improvements over the baseline (0.5106 versus

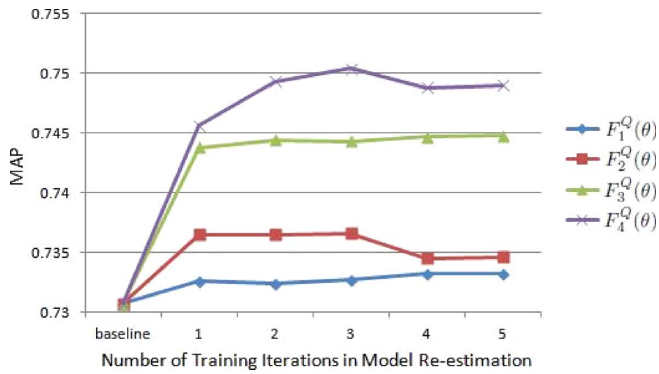

Fig. 5. Experimental results of Lecture with different objective functions and different number of training iterations in acoustic model re-estimation when the initial acoustic models were θ_{adp2} and N = 5 (the relevance information of the top five segments were given).

TABLE VI
EXPERIMENTAL RESULTS OF **NEWS** FOR MODEL-BASED SHORT-TERM CONTEXT USER RELEVANCE FEEDBACK WITH OBJECTIVE FUNCTIONS $F_4^Q(\theta)$ FOR N =5,10,15,20. ACOUSTIC MODEL RE-ESTIMATION WAS STARTED WITH THE ACOUSTIC MODELS $\theta_{basic-n}$ USED FOR GENERATING THE LATTICES. THE SUPERSCRIPT LABEL $^{(0)}$ INDICATES SIGNIFICANTLY BETTER THAN THE BASELINE

	baseline	Number of Feedback Segments (N)			
		N=5	**N=10**	**N=15**	**N=20**
MAP	0.6302	$0.6464^{(0)}$	$0.6480^{(0)}$	$0.6482^{(0)}$	$0.6405^{(0)}$

TABLE VII
EXPERIMENTAL RESULTS OF **LECTURE** FOR MODEL-BASED LONG-TERM CONTEXT USER RELEVANCE FEEDBACK WITH DIFFERENT NUMBERS OF TRAINING QUERIES FOR $N = 5$ (RELEVANCE INFORMATION FOR TOP FIVE SEGMENTS WERE GIVEN). ACOUSTIC MODEL RE-ESTIMATION CAN BE STARTED WITH ACOUSTIC MODELS $\theta_{basic-l}$, θ_{adp1}, OR θ_{adp2}, AND THE BASELINE MAPS WITHOUT RELEVANCE FEEDBACK ARE 0.4819, 0.6189, AND 0.7307 FOR LATTICES GENERATED BY $\theta_{basic-l}$, θ_{adp1}, AND θ_{adp2}, RESPECTIVELY. THE SUPERSCRIPT LABELS $^{(0)}$ INDICATE SIGNIFICANTLY BETTER THAN THE BASELINE

Cross Validation	Number of Training queries	Acoustic Unit Coverage	Initial Acoustic Models		
			$\theta_{basic-l}$	θ_{adp1}	θ_{adp2}
baseline	-	-	0.4819	0.6189	0.7307
2-fold	40	37%	$0.4999^{(0)}$	$0.6304^{(0)}$	0.7401
4-fold	60	46%	$0.5021^{(0)}$	$0.6386^{(0)}$	0.7410
8-fold	70	47%	$0.5087^{(0)}$	$0.6400^{(0)}$	$0.7444^{(0)}$
16-fold	75	50%	$0.5099^{(0)}$	$0.6419^{(0)}$	$0.7459^{(0)}$

0.4819 for $\theta_{basic-l}$, 0.6416 versus 0.6189 for θ_{adp1}, and 0.7504 versus 0.7307 for θ_{adp2}).

Fig. 5 shows the results of **Lecture** with different objective functions and different numbers of training iterations when the initial acoustic models were θ_{adp2} and $N = 5$. The results with three iterations in Fig. 5 are exactly those listed in a row of Table V. Based on Fig. 5 we observed that the results of model re-estimation converged in only a few iterations. For other cases in Table V similar phenomena were also observed; Fig. 5 is a typical example. Such results indicate the concept proposed here is practically feasible since the training can be completed quickly online. $F_4^Q(\theta)$ with three training iterations were used for the model-based approach in the following experiments.

Table VI shows the experimental results of **News** for the objective function $F_4^Q(\theta)$ in (13) with different N (N =5,10,15,20). Similar to Table V, significant improvements over the baseline were observed.

2) Long-Term Context User Relevance Feedback (Case (ϵ) in Table I): In long-term context user relevance feedback experiments, the 80 queries for **Lecture** and 160 queries for **News** were separated into 2, 4, 8, or 16 folds for cross validation. Each fold was selected once as the testing query set with the other folds set aside as the training query set. For all training queries, we assume the relevance information has been given for top five segments ($N = 5$) in the first-pass returned lists, and we applied $F_5^{lt}(\theta)$ in (19) to train a new set of acoustic models using the objective function $F_4^Q(\theta)$. The new acoustic models were used to rescore all the lattices in the spoken archive.

Table VII lists the experimental results of **Lecture** with different numbers of training queries. In each test for 2-, 4-, 8-, or 16-fold cross validation respectively 40, 20, 10, or 5 queries were tested, and 40, 60, 70, or 75 queries were used in training. Clearly, the number of training queries affects the performance of the re-estimated acoustic models. In addition, if the acoustic units[2] of a new query do not exist in the training query set, the retrieval performance of the new query may not be influenced by the long-term context relevance feedback; hence, the percentage of the acoustic units shared by the training and testing query sets may play an even greater role in the performance of long-term context relevance feedback.

The acoustic unit coverage listed in Table VII is the averaged percentage of triphones appearing in the test queries that also appear in the training query set. In Table VII, we started acoustic model re-estimation with acoustic models $\theta_{basic-l}$, θ_{adp1}, or θ_{adp2}, and the new acoustic models were used to rescore the lattices generated by the initial acoustic models. Although MAP improvements in general increased with the number of training queries, results showed that it is possible to obtain significant improvements with only 40 training queries each with five labeled segments with initial models $\theta_{basic-l}$ or θ_{adp1}.

[2]triphone models for **Lecture** and initial plus final models for **News**

TABLE VIII
EXPERIMENTAL RESULTS OF **NEWS** FOR MODEL-BASED LONG-TERM CONTEXT USER RELEVANCE FEEDBACK WITH DIFFERENT NUMBERS OF TRAINING QUERIES FOR $N = 5$ (RELEVANCE INFORMATION FOR TOP FIVE SEGMENTS WERE GIVEN). THE MODEL RE-ESTIMATION WAS STARTED FROM THE ACOUSTIC MODELS $\theta_{basic-n}$ USED TO GENERATE THE LATTICES. THE SUPERSCRIPT LABELS $^{(0)}$ INDICATE SIGNIFICANTLY BETTER THAN THE BASELINE

Cross Validation	Number of Training queries	Acoustic Unit Coverage	MAP
baseline	-	-	0.6302
2-fold	80	87%	0.6319
4-fold	120	93%	$0.6340^{(0)}$
8-fold	140	95%	$0.6361^{(0)}$
16-fold	150	96%	$0.6362^{(0)}$

Table VIII lists the experimental results of **News** with different numbers of training queries. In each test for 2-, 4-, 8-, or 16-fold cross validation respectively 80, 40, 20, or 10 queries were tested, while 80, 120, 140, or 150 queries were used in training. Acoustic unit coverage in Table VIII is the averaged percentage of initial and final models appearing in the test queries that also appear in the training query set. The acoustic unit coverage of **News** in Table VIII is much higher than **Lecture** in Table VII as there were 4602 triphones for **Lecture** but only 147 initial plus final models for **News**. The acoustic models were re-estimated from $\theta_{basic-n}$ which was used to generating the lattices. Although the training and testing query set share many initial and final models for **News**, the improvements of **News** in Table VIII was not as large as **Lecture** in Table VII probably because the initial acoustic models $\theta_{basic-n}$ were trained on broadcast news which is already relatively matched to the target archive. However, the experiment results showed that with more than 120 training queries significant improvements were obtained.

IX. INTEGRATION OF DIFFERENT RELEVANCE FEEDBACK SCENARIOS AND APPROACHES

Above we discussed results from experiments on individual relevance feedback scenarios and approaches as summarized

TABLE IX
EXPERIMENTAL MAP RESULTS FOR THE INTEGRATED FRAMEWORK (FIG. 4). **LECTURE** AND **NEWS** CORRESPOND TO THE TWO SPOKEN ARCHIVES. THE THREE COLUMNS FOR **LECTURE** REPRESENT THE RESULTS FROM LATTICES GENERATED USING THE THREE SETS OF ACOUSTIC MODELS $\theta_{basic-l}$, θ_{adp1}, AND θ_{adp2}; LIKEWISE THE $\theta_{basic-n}$ MODEL FOR THE **NEWS** COLUMN. THE ROWS (a) THROUGH (e) ARE THE RESULTS FOR THE LISTS (a) THROUGH (e) SHOWN IN FIG. 4. THE UPPER AND LOWER SECTIONS ARE THE RESULTS WITHOUT AND WITH LONG-TERM CONTEXT USER RELEVANCE FEEDBACK. THE SUPERSCRIPT LABELS $^{(a)}$, $^{(b)}$, $^{(c)}$, AND $^{(d)}$ RESPECTIVELY INDICATE SIGNIFICANTLY BETTER THAN LISTS (a), (b), (c), AND (d) IN THE SAME PART AND THE SAME COLUMN

long-term context user relevance feedback	list	cases applied	Different Spoken Archives and Initial Acoustic Models			
			Lecture			**News**
			$\theta_{basic-l}$	θ_{adp1}	θ_{adp2}	$\theta_{basic-n}$
NO	(a)	None	0.4819	0.6189	0.7307	0.6302
	(b)	(β)	$0.5208^{(a)}$	$0.6477^{(a)}$	$0.7423^{(a)}$	$0.6381^{(a)}$
	(c)	(β)+(α)	$0.5371^{(a)(b)}$	$0.6636^{(a)(b)}$	$0.7465^{(a)}$	$0.6443^{(a)(b)}$
	(d)	(β)+(γ)	$0.5321^{(a)}$	$0.6592^{(a)}$	$0.7544^{(a)(b)}$	$0.6522^{(a)(b)(c)}$
	(e)	(β)+(α)+(γ)	$0.5374^{(a)(b)}$	$0.6671^{(a)(b)(d)}$	$0.7599^{(a)(b)(c)}$	$0.6528^{(a)(b)(c)}$
YES	(a)	(ϵ)	0.5058	0.6352	0.7411	0.6330
	(b)	(ϵ)+(β)	$0.5368^{(a)}$	$0.6517^{(a)}$	$0.7513^{(a)}$	$0.6403^{(a)}$
	(c)	(ϵ)+(β)+(α)	$0.5429^{(a)(b)}$	$0.6719^{(a)(b)(d)}$	$0.7601^{(a)(b)}$	$0.6454^{(a)(b)}$
	(d)	(ϵ)+(β)+(γ)	$0.5569^{(a)(b)(c)}$	$0.6685^{(a)(b)}$	$0.7557^{(a)}$	$0.6615^{(a)(b)(c)}$
	(e)	(ϵ)+(β)+(α)+(γ)	$0.5601^{(a)(b)(c)}$	$0.6751^{(a)(b)(d)}$	$0.7640^{(a)(b)(d)}$	$0.6629^{(a)(b)(c)(d)}$

in Table I; here, we present and discuss results of experiments using the integrated framework described in Section VI-B and shown in Fig. 4.

The results of **Lecture** and **News** are listed in Table IX. The upper part and lower part correspond to experiments without and with long-term context user relevance feedback. In the upper part, rows (a) to (e) are the MAP values for the five lists (a) to (e) shown in Fig. 4, which are ranked by the relevance score functions listed in Table II without long-term context user relevance feedback. The three columns of **Lecture** here correspond to the results from lattices generated by the acoustic models $\theta_{basic-l}$, θ_{adp1}, or θ_{adp2}; likewise, the rightmost column is for **News** results generated using $\theta_{basic-n}$. That is, the acoustic models θ used to evaluate the relevance score functions in Table II were $\theta_{basic-l}$, θ_{adp1}, or θ_{adp2} for **Lecture** and $\theta_{basic-n}$ for **News**. In the upper part of Table IX, row (a) is the baseline, row (b) is for case (β), and rows (c), (d), and (e) are for cases (α), (γ), and (α)+(γ) applied on top of case (β).

In the lower part of Table IX, the long-term context user relevance feedback with 2-fold cross validation was then applied in addition. In 2-fold cross validation for long-term context user relevance feedback, 40 training queries for **Lecture** and 80 for **News** first went through the pseudo-relevance feedback of the upper part [case (β)], and then the acoustic model parameters were updated using the user relevance information applied on the results (row (b), upper part) thus obtained from these training queries. These updated acoustic models were then used in the experiments on the other queries presented as rows (a) through (e) in the lower part. Hence, for these tests, case (ϵ) was actually performed in addition in the beginning, in contrast to the upper part of the table. Throughout Table IX the superscript labels $^{(a)}$, $^{(b)}$, $^{(c)}$, and $^{(d)}$, respectively, indicate significantly better than rows (a), (b), (c), and (d) in the corresponding part and column.

A. No Long-Term Context User Relevance Feedback – Upper Part of Table IX

Here we discuss the upper part of Table IX. The improvements of pseudo-relevance feedback were significant [row (b) versus row (a)] under all the conditions ($\theta_{basic-l}$, θ_{adp1}, and θ_{adp2} for **Lecture**, and $\theta_{basic-n}$ for **News**) as already shown in Section VIII-A2 (row (b) corresponds to $M = 9$ for **Lecture** and $M = 7$ for **News** in Table IV). After pseudo-relevance feedback, when the user browsed through the retrieval results [list (b)], the relevance information of the top five segments was provided by the user. This information was first used to improve the retrieval results for the current query, or for the short-term context. Note that this is different from the previous experiments in Sections VIII-A2 and VIII-B1, in which the user provided relevance information for the top five segments of the first-pass baseline results [list (a)]. Here the information obtained was instead for the top five segments on list (b), because list (b) is the results the user actually browsed and clicked through when using the system. This is cases (α) and/or (γ) applied on top of case (β); the results are in rows (c), (d), and (e) in the upper part of Table IX. As we can see from the table, although pseudo-relevance feedback already leads to significant improvements, the results can be further improved with user relevance feedback in the short-term context, for either example-based [row (c)] or model-based [row (d)] approaches or both [row (e)] in short-term context, as compared to the pseudo-relevance feedback [row (b)].

From the upper part of Table IX, it seems that example-based and model-based relevance feedback [list (c) versus list (d)] were comparable for **Lecture**, but for **News**, model-based relevance feedback significantly outperformed example-based. This is probably because **News** contains the speech of many different speakers with high degree of speaker variations; example-based approaches that depend on DTW-derived acoustic feature similarities obtained without distribution modeling yield poor performance. In contrast, model-based approaches directly adjust model parameters and thus are able to better model speaker variations. Because example-based and model-based approaches use the relevance information in different ways, it is reasonable that their effects may be additive. This is verified in the next row (e), where the integration surpassed each individual approach under all conditions, and the results of integration are now significantly better than that

for pseudo-relevance feedback in all the cases [row (e) versus row (b)].

B. With Long-Term Context User Relevance Feedback – Lower Part of Table IX

Now consider the lower part of Table IX for long-term context user relevance feedback. As mentioned before, the relevance information for the top five spoken segments was collected when the user browsed the results after pseudo-relevance feedback (list (b) in Fig. 4 and row (b) in the upper part). All such relevance information for the 40 training queries for **Lecture** (or 80 for **News**) were used for model training in long-term context. Since the top five segments were labeled for each query among the training queries, the system collected the relevance information of 200 segments for **Lecture** and 400 segments for **News** in each trial when training. The parameters of the acoustic models previously used to generate the lattices were updated using the relevance information of these segments, and these new models were used to rescore the lattices of the whole spoken archive. After rescoring, the other 40 testing queries for **Lecture** (or 80 for **News**) underwent the same processes including pseudo-relevance feedback and user relevance feedback in short-term context, yielding the results in rows (a) to (e) after long-term context feedback in the lower part of the table.

In the lower part of the table, row (a) is the results ranked based on $S(Q, X|\theta^*_{lt})$, where θ^*_{lt} was re-estimated from $\theta_{basic-l}$, θ_{adp1}, or θ_{adp2} by the relevance information of the 200 segments for **Lecture**, or from $\theta_{basic-n}$ by the relevance information of the 400 segments of **News**. We found that regardless of the initial acoustic models, the relevance feedback in long-term context always yielded improvement (compare rows (a) in the upper and lower parts). After this long-term context feedback, pseudo-relevance feedback and example-based and model-based methods in short-term context were applied again in addition, yielding the results in rows (b) to (e) in the lower part. Hence, rows (a) through (e) in the lower part are in parallel with those in the upper part, except that they are performed on top of case (ϵ). We can see in most cases the effects of these approaches were still additive, yielding incremental performance improvements (rows (b) to (e) in the lower part). Comparing respectively lists (b) through (e) in lower part to lists (b) through (e) in upper part, we found that with long-term context feedback the results of pseudo-relevance feedback and short-term context user relevance feedback were all improved.

Finally, in Table IX, a comparison of the last row of column $\theta_{basic-l}$ (0.5601) and first row of column θ_{adp1} (0.6189), or the last row of column θ_{adp1} (0.6751) and the first row of column θ_{adp2} (0.7307), seems to suggest that speaker adaptation techniques yield greater performance improvements than relevance feedback. However, the speaker adaptation applied here used a large quantity of adaptation data (more than 20 minutes, very difficult to obtain in practice), whereas relevance feedback used the relevance information of only five segments for each query. Practically it would be much simpler for the user to click five segments when browsing through the results than it would be to label 20 minutes of speech. Moreover, these results show that the proposed relevance feedback approaches can be integrated with speaker adaptation techniques if adaptation data are available, yielding additive improvements to performance. For example, with speaker adaptation (θ_{adp2}), the MAP improved from 0.7307 (first row of θ_{adp2}) to 0.7640 (last row of θ_{adp2}) with the proposed approaches.

X. Conclusion

We presented example-based and model-based approaches for STD relevance feedback. Example-based approaches that take into account acoustic feature similarities with detected examples were shown effective in pseudo-relevance feedback and short-term context user relevance feedback scenarios. For model-based approaches, where acoustic model parameters are adjusting according to the results of relevance feedback, the best performance was obtained by using objective functions that take into account the nature of the retrieval task and the unlabeled segments; results were shown for both short- and long-term context relevance feedback. Relevance feedback scenarios and approaches were shown to yield improved performance individually, and yielded still greater performance improvements when integrated together.

References

[1] L.-S. Lee and B.-L. Chen, "Spoken document understanding and organization," *IEEE Signal Process. Mag.*, vol. 22, no. 5, pp. 42–60, Sep. 2005.

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

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

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

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

[6] H.-L. C. Y.-C. Pan 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 *Proc. ASRU*, 2007.

[7] T. Hori, I. Hetherington, T. Hazen, and J. Glass, "Open vocabulary spoken utterance retrieval using confusion networks," in *Proc. ICASSP*, 2007, pp. 73–76.

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

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

[10] B. Logan, J.-M. Van Thong, and P. Moreno, "Approaches to reduce the effects of OOV queries on indexed spoken audio," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 899–906, Oct. 2005.

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

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

[13] D. Wang, J. Frankel, J. Tejedor, and S. King, "A comparison of phone and grapheme-based spoken term detection," in *Proc. ICASSP*, 2008, pp. 4969–4972.

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

[15] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *Proc. ICASSP*, 2006, pp. 949–952.

[16] S. w. Lee, K. Tanaka, and Y. Itoh, "Combining multiple subword representations for open-vocabulary spoken document retrieval," in *Proc. ICASSP*, 2005, pp. 505–508.

[17] S. Meng, P. Yu, J. Liu, and F. Seide, "Fusing multiple systems into a compact lattice index for Chinese spoken term detection," in *ICASSP*, 2008, pp. 4345–4348.

[18] Y. c. Pan, H. l. Chang, and L. s. Lee, "Type-II dialogue systems for information access from unstructured knowledge sources," in *Proc. ASRU*, 2007.

[19] S.-Y. Kong, M.-R. Wu, C.-K. Lin, Y.-S. Fu, and L.-S. Lee, "Learning on demand – Course lecture distillation by information extraction," in *Proc. ICASSP*, 2009, pp. 4709–4712.

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

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

[22] 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 *Proc. ICASSP*, 2009, pp. 4873–4876.

[23] R. Wallace, R. Vogt, B. Baker, and S. Sridharan, "Optimising figure of merit for phonetic spoken term detection," in *Proc. ICASSP*, 2010, pp. 5298–5301.

[24] S. Srinivasan and D. Petkovic, "Phonetic confusion matrix based spoken document retrieval," in *Proc. SIGIR*, 2000.

[25] H. Nanjo and T. Kawahara, "A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding," in *Proc. ICASSP*, 2005, pp. 1053–1056.

[26] T. Shichiri, H. Nanjo, and T. Yoshimi, "Minimum Bayes-risk decoding with presumed word significance for speech based information retrieval," in *Proc. ICASSP*, 2008, pp. 1557–1560.

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

[28] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 257–265, May 1997.

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

[30] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART System Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Printice-Hall, 1971.

[31] I. Ruthven and M. Lalmas, "A survey on the use of relevance feedback for information access systems," *Knowl. Eng. Rev.*, pp. 95–145, 2003.

[32] S. E. Robertson and K. S. Jones, "Relevance weighting of search terms," *J. Amer. Soc. Inf. Sci.*, pp. 129–146, 1976.

[33] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. 10th Int. Conf. Inf. Knowl. Manage. CIKM '01*, New York, 2001, pp. 403–410.

[34] P. Hong, Q. Tian, and T. Huang, "Incorporate support vector machines to content-based image retrieval with relevance feedback," in *Proc. Int. Conf. Image Process.*, 2000.

[35] S. MacArthur, C. Brodley, and C.-R. Shyu, "Relevance feedback decision trees in content-based image retrieval," in *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*, 2000.

[36] J. Xin and J. Jin, "Learning from user feedback for image retrieval," in *Proc. Joint Conf. 4th Int. Conf. Inf. Commun. Signal Process. and 4th Pacific Rim Conf. Multimedia*, 2003.

[37] N. Vasconcelos and A. Lippman, "Bayesian relevance feedback for content-based image retrieval," in *Proc. IEEE Workshop Content-Based Access of Image and Video Libraries*, 2000.

[38] R. Yan, A. Hauptmann, and R. Jin, "Negative pseudo-relevance feedback in content-based video retrieval," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003.

[39] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. CIVR*, 2003.

[40] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, 2007.

[41] 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 *Proc. Interspeech*, 2010.

[42] H.-Y. Lee and L.-S. Lee, "Integrating recognition and retrieval with user feedback: A new framework for spoken term detection," in *Proc. ICASSP*, 2010, pp. 5290–5293.

[43] C.-P. Chen, H.-Y. Lee, C.-F. Yeh, and L.-S. Lee, "Improved spoken term detection by feature space pseudo-relevance feedback," in *Proc. Interspeech*, 2010.

[44] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," in *Proc. IEEE Workshop Autom. Speech Recogn. Understanding ASRU'09*, 2009.

[45] H.-Y. Lee, C.-P. Chen, C.-F. Yeh, and L.-S. Lee, "A framework integrating different relevance feedback scenarios and approaches for spoken term detection," in *Proc. SLT*, 2010.

[46] D. Kelly and J. Teevan, *SIGIR Forum*, vol. 37, pp. 18–28, 2003.

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

[48] D. Kelly and N. J. Belkin, "Display time as implicit feedback: Understanding task effects," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.*, 2004.

[49] X. Shen, B. Tan, and C. Zhia, "Context sensitive information retrieval using implicit feedback," in *Proc. SIGIR*, 2005.

[50] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *Multimedia Syst.*, vol. 8, pp. 536–544, 2003.

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

[52] S. Yu, D. Cai, J.-R. Wen, and W.-Y. Ma, "Improving pseudo-relevance feedback in web information retrieval using web page segmentation," in *Proc. 12th Int. Conf. World Wide Web*, 2003.

[53] J. Xu and W. B. Croft, "Query expansion using local and global document analysis," in *Proc. 19th Annu. Int. ACM SIGIR Conf. Research Develop. Inf. Retrieval*, 1996.

[54] V. Lavrenko and W. B. Croft, "Relevance based language models," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.*, 2001.

[55] O. Kurland, L. Lee, and C. Domshlak, "Better than the real thing?: Iterative pseudo-query processing using cluster-based language models," in *Proc. 28th Ann. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2005.

[56] T. Tao and C. Zhai, "Regularized estimation of mixture models for robust pseudo-relevance feedback," in *Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.* , 2006.

[57] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting good expansion terms for pseudo-relevance feedback," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008.

[58] K. S. Lee, W. B. Croft, and J. Allan, "A cluster-based resampling method for pseudo-relevance feedback," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2008.

[59] Y. Lv and C. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proc. 18th ACM Conf. Inf. Knowl. Manage.*, 2009.

[60] Y. Lv and C. Zhai, "Positional relevance model for pseudo-relevance feedback," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2010.

[61] W.-H. Lin, R. Jin, and A. Hauptmann, "Web image retrieval re-ranking with relevance model," in *Proc. IEEE/WIC Int. Conf. Web Intell. WI '03.* , 2003, pp. 242–248.

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

[63] 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 *Proc. Interspeech*, 2007.

[64] J. Mamou, B. Ramabhadran, and O. Siohan, "Vocabulary independent spoken term detection," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retriev.*, 2007.

[65] P. Yu, K. Chen, L. Lu, and F. Seide, "Searching the audio notebook: Keyword search in recorded conversations," in *Proc. Conf. Human Lang. Technol. Empir. Methods Natural Lang. Process.*, 2005.

[66] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguist.*, 2005.

[67] S. Parlak and M. Saraclar, "Spoken term detection for Turkish broadcast news," in *Proc. ICASSP*, 2008, pp. 5244–5247.

[68] Y. Zhang and J. Glass, "Towards multi-speaker unsupervised speech pattern discovery," in *Proc. ICASSP*, 2010, pp. 4366–4369.

[69] Y. Zhang and J. Glass, "Unsupervised spoken keyword spotting via segmental Dtw on Gaussian posteriorgrams," in *Proc. ASRU*, 2009.

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

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

[72] J.-T. Chien and M.-S. Wu, "Minimum rank error language modeling," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 267–276, Feb. 2009.
[73] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2003.
[74] B.-Y. Liang, "Acoustic models for continuous mandarin speech recognition," M.S. thesis, National Taiwan Univ., Taipei, Taiwan, 1998.

Hung-yi Lee was born in 1986. He received the B.S. and M.S. degrees in electronic engineering and communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2008 and 2010, respectively. He is currently pursuing the Ph.D. degree in the Department of Communication Engineering, National Taiwan University, Taipei, Taiwan.

His research focuses on spoken content retrieval.

Chia-ping Chen was born in 1987. She received the B.S. degree in communication engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 2009, and the M.S. degree in communication engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2011.

She is currently an Engineer with MediaTek, Inc., Hsinchu. Her research has been focused on spoken content retrieval.

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

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

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