Improved Spoken Term Detection by Discriminative Training of Acoustic Models based on User Relevance Feedback

Hung-yi Lee, Chia-ping Chen, Ching-feng Yeh, Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
{tlkagkb93901106,coward7652}@yahoo.com.tw, r98942056@ntu.edu.tw, lslee@gate.sinica.edu.tw

Abstract

In a previous paper [1], we proposed a new framework for spoken term detection by exploiting user relevance feedback information to estimate better acoustic model parameters to be used in rescoring the spoken segments. In this way, the acoustic models can be trained with a criterion of better retrieval performance, and the retrieval performance can be less dependent on the existence of a set of acoustic models well matched to the corpora to be retrieved. In this paper, a new set of objective functions for acoustic model training in the above framework was proposed considering the nature of retrieval process and its performance measure, and discriminative training algorithms maximizing the objective functions were developed. Significant performance improvements were obtained in preliminary experiments.

Index Terms: spoken term detection, user relevance feedback

1. Introduction

Spoken term detection is to return a list of spoken segments containing the term requested by the user selected from a large spoken document archive. It is a key technology for voice-based information retrieval, and believed to be very important in the future when people try to access the multimedia contents based on its audio signals. In general, there are two stages for spoken term detection as shown in Fig. 1. In the first stage, the audio content is transformed into 1-best transcriptions or lattices by a recognition engine based on a set of acoustic models. Lattices are preferred since they include multiple recognition hypotheses, especially when the accuracy in 1-best transcriptions is relatively low [2]. In the second stage, the retrieval engine identifies and returns to the user a list of relevant spoken segments based on the user query. Some approaches such as a confusion matrix may be used here in the retrieval model to better handle the recognition errors within the lattices.

If the recognition engine can transcribe the audio signals into text perfectly, the spoken term detection problem is reduced to a text retrieval task. However, spoken term detection is very hard because it is practically very difficult to obtain acoustic and language models robust enough for the huge quantities of target audio corpora over the Internet generated by different speakers under different conditions with varying environments. Although many very efficient approaches [3, 4, 5] have been proposed to enhance the detection performance due to the relatively poor recognition output, the training data reasonably matched to the audio corpus to be retrieved is often necessary, which seriously limits the scope of practical applications of spoken term detection. In a previous work [1], we introduced the user relevance feedback in Fig. 1 to construct a loop, with which the two modules of recognition and retrieval can be jointly improved.

In other words, when a query is entered by a user, the retrieval system offers a ranked list of matched spoken segments to the user. If the user gives some feedback to the system, for example he/she selects items 1 and 3 shown in Fig. 1 as relevant but item 2 as irrelevant, a new set of acoustic models can then be estimated based on the feedback and so on. This approach is helpful to alleviate the need for reasonably matched audio data for training the acoustic models, and can be complementary to and integrated with many approaches recently proposed to compensate for the relatively poor recognition output.

In this paper, a whole set of improved techniques for the above approach of acoustic model training based on user relevance feedback to achieve better spoken term detection performance is proposed. This is quite different from the conventional acoustic model training approaches designed for speech recognition:

1. The system input includes only whether a spoken segment is relevant to a query or not, not the transcription of any utterance [6].

2. The goal is to improve the retrieval performance rather than the recognition accuracy.

The proposed techniques include a new set of objective functions properly considering the nature of retrieval process and its performance measure, and discriminative training algorithms optimizing these objective functions were developed. Significant performance improvements were obtained.

2. Scenarios for User Relevance Feedback

Although all users are reluctant to give relevance feedback explicitly, implicit feedback is practically very reasonable which means the system analyzes user’s behaviour on-line to get the feedback information, while the user doesn’t know he/she is in a feedback procedure. One example approach is the process of click-through data [7]. Suppose the transcription is displayed beside each spoken segment on the returned list, if a user clicks the third spoken segment on the list without clicking the first
two, it is reasonable for the system to assume the third segment to be relevant, but the first two are irrelevant.

Two scenarios for user relevance feedback are considered in this paper. In the first scenario, after a query is entered, the retrieval system performs the first pass retrieval and returns a list of spoken segments. Some segments in the returned list are then labelled as relevant or irrelevant by the user, for example by the click-through data process. Those labelled segments are taken as training data, and those unlabelled as testing data. The parameters of a new set of acoustic models are then estimated on-line according to the feedback information. The first pass returned list is finally reranked based on the new set of acoustic models. In this scenario, the acoustic models are trained based on the feedback information for the query entered, with a goal to improve the retrieval performance for the specific query. Because the training data is very limited, it is possible to perform the acoustic model training and reranking of returned list on-line in real time when the user clicks through the returned list. This scenario is practically reasonable, and is referred to as short-term context feedback in the literature of text retrieval [8].

In the second scenario, the system collects the feedback information of a set of queries entered in a certain period of time to estimate a new set of acoustic models. Then the lattices of the whole corpus to be retrieved are rescored based on the new set of acoustic models with a goal to improve the retrieval performance for queries to be entered in the future. The past queries are taken as training queries, and the new queries as testing queries. This scenario is referred to as long-term context feedback [8].

3. Acoustic Model Training

The definition of the relevance score is in Section 3.1, and the objective functions and optimization algorithms are in Sections 3.2 and 3.3.

3.1. Relevance Score

An observation sequence $X$ for a spoken segment in the audio corpus to be retrieved is first transcribed into a lattice, which is a weighted directed acyclic graph (DAG) $\{N, A\}$, where $N$ is the set of nodes, and $A$ is the set of arcs. Let $\text{word}(a)$ denotes the word hypotheses of arc $a$. When a query $Q$ is entered, all observation sequence $X$ in the corpus to be retrieved are ranked based on the relevance score $S(Q, X|\theta)$ with respect to $\theta$ where

$$S(Q, X|\theta) = \sum_{a \in A, \text{word}(a) = Q} P(a|X, \theta),$$

(1)

which is the sum of posterior probabilities of all words in the corresponding lattice of $X$, where $\theta$ is the set of acoustic model parameters. Similar relevance score is widely used in many spoken term detection techniques [9]. The posterior probability of an arc $a$ for an observation sequence $X$ is defined as

$$P(a|X, \theta) = \frac{\sum_{u \in W_a} P_0(X|u)P(u)}{\sum_{u \in W_a} P_0(X|u)P(u)},$$

(2)

which is the sum of the likelihood of all word sequences $u$ in the lattice containing arc $a$ divided by the sum of likelihood for all possible word sequences $W_a$ in the corresponding lattice of $X$. $P_0(X|u)$ is the likelihood for the observation sequence $X$ given the word sequence $u$ with the acoustic model $\theta$. Here we assume the query $Q$ is a single word through this paper for simplicity, although extension to longer queries is trivial.

3.2. Objective Function

After having some positive (relevant) and negative (irrelevant) examples for a certain query $Q$ from user feedback, the system estimates a set of new acoustic models by maximizing an objective function. The likelihood $P_0(X|u)$ of the arcs in the lattices corresponding to the spoken segments in the returned list is recomputed based on the new acoustic models, and then the relevance score $S(Q, X|\theta)$ of each spoken segment is modified accordingly. The returned list is finally reranked based on the new relevance scores.

The first objective function $F_1$ (Eq (3)) to be maximized is the sum of relevance scores of all positive examples,

$$F_1^Q(\theta) = \sum_{X^Q} S(Q, X^Q|\theta),$$

(3)

where $X^Q$ is the observation sequence for a positive example with respect to the query $Q$. Objective function $F_2$ (Eq (4)) is then the sum of the distances between all positive and negative example pairs,

$$F_2^Q(\theta) = \sum_{X^Q, X^Q} |S(Q, X^Q|\theta) - S(Q, X^Q|\theta)|,$$

(4)

where $X^Q$ is the observation sequence for a negative example with respect to the query $Q$.

However, we have to note that maximizing the distances between all pairs of positive/negative examples do not necessarily contribute to the improvement in the the evaluation measure of the retrieval process. In retrieval, it is usually preferred that all relevant documents are ranked higher than all irrelevant documents. This implies the relative positions of all positive examples with respect to all negative examples are more important than their individual absolute relevance scores. Most popular evaluation measures of retrieval, such as mean average precision (MAP) [10], are designed in this way. On the one hand, an example pair containing a negative example having relevance score higher than a positive example is definitely undesired because it degrades the retrieval performance. On the other hand, if a positive example already has higher relevance score than all negative examples, the increase of the relevance score of this positive example cannot benefit the retrieval performance any more. With these considerations, we have a new objective function $F_3$ (Eq (5)) integrating the ranking requirements for retrieval purposes.

$$F_3^Q(\theta) = \sum_{X^Q, X^Q} \{S(Q, X^Q|\theta) - S(Q, X^Q|\theta)\} \delta(X^Q, X^Q),$$

(5)

where

$$\delta(X^Q, X^Q) = \begin{cases} 1 & S(X^Q, Q|\theta) < S(X^Q, Q|\theta) \\ 0 & \text{otherwise} \end{cases}.$$
of acoustic models which keeps the positive examples ranked at the top of the returned list to prevent the above overfitting. This can be achieved by modifying the objective function \( F_3 \) (Eq (5)) into \( F_4 \) (Eq (7)),

\[
F_4^Q(\theta) = F_3^Q(\theta) + \rho \sum_{X^Q_{i,n} \in X^Q_{i,n}} [S(Q, X^Q_{i,n}) - S(Q, X^Q_{i,n}')] ,
\]

where \( X^Q_{i,n} \) is an unlabelled segment within the returned list, and we set \( \rho = 0.1 \) in our experiment. Eq (7) can be viewed as a smoothing method keeping the unlabelled segments to be located relatively lower than positive examples.

All the above are for short-term context feedback. For long-term context feedback, we collect a set of queries with user relevance feedback, and the system tries to estimate a new set of acoustic models with improved retrieval performance for all the training queries, and expect the improvements can be generalized to other queries not in the training query set. This can be easily achieved by replacing \( F_3, F_4 \) in Eqs (5), (7) by \( F_3' \) and \( F_4' \) (Eq (8)) when we have a training query set,

\[
F_3'(\theta) = \sum_Q F_3^Q(\theta), \quad F_4'(\theta) = \sum_Q F_4^Q(\theta), \quad \text{where the summation is over all training queries } Q .
\]

### 3.3. Optimization

All the objective functions presented in Section 3.2 can be optimized using the weak-sense auxiliary function in the same way as the minimum phone error (MPE) discriminative training [11]. The objective function of MPE maximizes the expected phone accuracy as in Eq (9).

\[
F_{MPE}(\theta) = \sum_{r=1}^{R} \frac{\sum_{u \in W_h(X_r)} P_h(X_r|u)P(u)A(u)}{\sum_{u \in W_h(X_r)} P_h(X_r|u)P(u)} ,
\]

where \( W_h(X_r) \) are the sets of all possible word sequences in the lattice for the utterance \( X_r \), \( R \) is the number of training utterances, and \( A(u) \) is the phone accuracy evaluated for the corresponding phone sequence of the word sequence \( u \). Taking \( F_2 \) in Eq (4) as an example, here we show all objective functions in Section 3.2 can be managed to have the same form as Eq (9) except defined for a word sequence \( u \) with different definition of \( A(u) \). By substituting Eq (2) into Eq (1), (1) can be written as Eq (10).

\[
S(Q, X|\theta) = \frac{\sum_{u \in W_h(X)} P_h(X|u)P(u)N(u, Q)}{\sum_{u \in W_h(X)} P_h(X|u)P(u)} ,
\]

where \( N(u, Q) \) is the number of occurrence of the word hypotheses \( \hat{Q} \) in the word sequence \( u \). Hence, by substituting Eq (10) into Eq (4), (4) can be written as Eq (11).

\[
F_2^Q(\theta) = \sum_{X^Q_r \in X^Q} \frac{\sum_{u \in W_h(X^Q_r)} P_h(X^Q_r|u)P(u)N(u, Q)}{\sum_{u \in W_h(X^Q_r)} P_h(X^Q_r|u)P(u)} + \sum_{X^Q_r \in X^Q} \frac{\sum_{u \in W_h(X^Q_r)} P_h(X^Q_r|u)P(u)N'u(u, Q)}{\sum_{u \in W_h(X^Q_r)} P_h(X^Q_r|u)P(u)}
\]

where \( W_h(X^Q_r), W_h(X^Q_r) \) are the sets of all possible word sequences in the lattices for the examples \( X^Q_r \) and \( X^Q_r \) respectively. \(|X^Q_r|, |X^Q_r|\) are the total number of positive and negative examples included in the evaluation in Eq (4) respectively, and \( N'(u, Q) \) is defined as \(-N(u, Q)\). Therefore, we can optimize Eq (11) in the same way as MPE by simply replacing \( A(u) \) in (9) by \(|X^Q_r|N(u, Q)\) and \(|X^Q_r'|N'(u, Q)\) as in Eq (11).

However, Eq (5) is not differentiable and can’t be optimized using the weak-sense auxiliary function straightforwardly because of the indicator function. To address this problem, we use an approximate function \( \delta(X^Q_r, X^Q_r) \) to replace \( \delta(X^Q_r, X^Q_r) \) at the \( i \)th training iteration,

\[
\delta(X^Q_r, X^Q_r') = \begin{cases} 1 & S(X^Q_r, Q|\theta^{t-1}) > S(X^Q_r, Q|\theta^t) \\ 0 & \text{otherwise} \end{cases}.
\]

Eq (12) uses the models obtained at the \( i \)-1th iteration to compute the indicator function used at the \( i \)-th iteration.

Because \( \theta^t \) is a constant, \( \delta(X^Q_r, X^Q_r) \) does not depend on the acoustic model parameters \( \theta \) to be optimized. By replacing \( \delta(X^Q_r, X^Q_r) \) by \( \delta(X^Q_r, X^Q_r) \), Eq (5) becomes differentiable with respect to \( \theta \). The optimization method of Eqs (3), (7) and (8) is then trivial.

### 4. Experiments

#### 4.1. Experimental Setup

The initial acoustic model was trained by Maximum Likelihood criterion with 4602 state-tied triphones spanned from 37 monophones using a corpus of noiseless read speech in Mandarin, which included 24.6 hours of data produced by 100 males and 100 females. 39-dimension MFCC was used as the feature. There were 5 states per triphone, and 24 mixtures per state. A lexicon with 10.7K words was used, and a tri-gram language model was trained with a 600M news data. We used 33 hours of recorded lectures for a course offered in National Taiwan University as the testing archive to be retrieved, which is quite noisy and spontaneous. So both the acoustic and language models are highly mismatched, yielding a relatively poor word accuracy of 50.26%. Each spoken segment in the corpus is transcribed into a lattice with beam width 50.

We modified the HTK tool, HMMRest, to implement the proposed approaches, and the parameters suggested by HTK Book were used in our experiments. The user relevance feedback proposed here was used to reestimate the acoustic model parameters including means, variances, transition probabilities, and mixture weights from the initial acoustic models which generated the lattices. MAP was used as our retrieval performance evaluation measure. 80 Chinese queries were manually selected, each with a single word. In short-term context feedback, we compared the MAP of the 80 queries before and after the reranking by the user relevance feedback. In long-term context feedback, the query set was separated into 16, 4 or 2 folds for cross validation. Each fold was selected once as the testing query set, while the others are the training query set.

#### 4.2. Experimental Results

##### 4.2.1. Short-term Context Feedback

For each query, we assumed the correct relevance information (positive or negative) of the top N (N = 5, 10, 15, 20) segments in the first-pass returned list was given. The segments below top N were then reranked based on the new acoustic models estimated with the feedback, while the positions of the top N segments in the ranking list were fixed. We compared the MAP score of the returned list before and after reranking. In this way,
the improvements in MAP scores were limited by the top N frozen segments since they dominated the MAP scores. This strategy also used previously [12] better reflects the user’s impression on the retrieval performance improvement, because the top N segments had been browsed by the user, and therefore re-arrangement of them does not make sense.

The experimental results of different objective functions ($F_1, F_2, F_3, F_4$) described in Section 3.2 with different N (N = 5, 10, 15, 20) are listed in Table 1. Because the data from user feedback is very limited, we were not able to have a validation set. Hence, a criterion is necessary to decide the number of acoustic model training iterations. Three criteria were tested here. “Oracle” denotes that we trained the acoustic model iteratively until the MAP score converged on the training query set. The results showed that it is possible to obtain significant improvements with only 60 training queries.

<table>
<thead>
<tr>
<th>No. of training queries</th>
<th>16-fold</th>
<th>4-fold</th>
<th>2-fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>phone coverage</td>
<td>50%</td>
<td>46%</td>
<td>37%</td>
</tr>
<tr>
<td>MAP Cri.</td>
<td>0.5141</td>
<td>0.5073</td>
<td>0.4977</td>
</tr>
<tr>
<td>Obj. Cri.</td>
<td>0.5141</td>
<td>0.5033</td>
<td>0.4947</td>
</tr>
</tbody>
</table>

Table 2: Experimental results of long-term context feedback with different numbers of training queries with different stopping criteria. The retrieval performance before rescoring (baseline) is 0.4819. The label \( \alpha \) indicates significantly better than baseline at the significance level 0.01 by pairwise t-test.

5. Concluding Remarks

In this paper, we propose improved approaches exploiting user relevance feedback to estimate acoustic model parameters for better performance in spoken term detection. Improved objective functions with corresponding optimization algorithms were developed, and significant performance improvements were obtained in preliminary experiments.

6. References