ABSTRACT

This paper presents a novel approach to spoken document retrieval where the speech recognition and information retrieval components are more tightly integrated. This is done by developing new recognizer and retrieval models where the interface between the two components is better matched and the component goals are consistent with the overall goal of the combined system. Experiments on radio news data show that the integrated approach can improve performance by 28% over a baseline system.

1. INTRODUCTION

Spoken document retrieval (SDR) is the task of automatically indexing and then retrieving relevant items from a large collection of recorded speech messages in response to a user specified natural language text query. In this paper, we propose a novel approach to SDR where the speech recognition and information retrieval components are more tightly integrated. This new approach represents a step towards moving away from the conventional method of having a cascade of independently operating recognition and retrieval components where the speech recognizer transforms the speech into text transcriptions which are then fed directly into a full-text retrieval system.

The cascade approach has the advantage of being modular in that different recognizers and retrieval models can be easily combined. However, there are some shortcomings. First, there is an input-output mismatch between the two components. The recognizer outputs errorful hypotheses while the retrieval model expects error-free text representations as input. Second, the two components have decoupled objectives. The two systems were originally designed to solve different problems and therefore have different objectives and make different assumptions. There is no guarantee that the goals of the components will be consistent with each other or with the overall goal of the combined system.

Speech recognition systems are usually designed to output the most likely symbol sequence (i.e., string of words or phones) corresponding to a given set of acoustic observations. High scoring alternative recognition hypotheses are typically not accounted for. The availability of additional hypotheses could be useful for retrieval since it offers the potential of including terms that would otherwise be missed.

Another issue is that recognizers are usually trained to try to minimize the error rate of the most likely symbol sequence. Although retrieval and recognition performance are correlated [5], it is not clear that minimizing the error rate is the best thing to do for retrieval purposes. One reason is that error rate is only computed using the single best recognition hypothesis; likely alternatives are not considered. Another is that all symbols are treated equally in computing the error rate. This means that in a word based system, function words are just as important as content words.

Information retrieval, all words are not created equal. In fact, the removal of commonly occurring function words (“stop words”) has been shown to improve retrieval performance [6].

Text-based retrieval systems are usually designed to index a collection of text documents and to perform term matching to find relevant documents in response to user-specified queries. Because the retrieval model is originally developed for use on text document collections where the words are assumed to be known with certainty, there is no explicit mechanism for dealing with errors in the document representations. Also, conventional text retrieval models generally do not make use of additional information that can be generated from the recognizer such as likelihood and confidence scores. These scores can be used to weight our belief in the accuracy of different recognition hypotheses which can be important when dealing with errorful transcriptions.

To develop a more integrated SDR approach, we create new recognizer and retrieval models where the interface between the two components is better matched and the goals of the two components are consistent with the overall goal of the combined system. First, we develop a novel probabilistic information retrieval model that makes direct use of information that can be computed by the speech recognizer. The retrieval model scores documents based on the relative change in the document likelihoods, expressed as the likelihood ratio of the conditional probability of the document given the query and the prior probability of the document before the query is specified. The likelihoods are computed using statistical language modeling techniques which eventually make use of p(\{P_D\}), the probability that term \( t \) occurs in spoken document \( D_t \). It is this probabilistic quantity that will serve as the interface between the two components. Second, we modify the objective of the speech recognizer to compute these \( p(\{P_D\}) \) probabilities given the speech waveform for spoken document \( D_t \) instead of finding the most likely symbol sequence given the acoustics observations. In this way, the interfaces of the speech recognition and retrieval components are better matched: the recognizer outputs term occurrence probabilities which the retrieval model expects as input. In addition, the goals of the two components are now consistent with the overall goal of the combined system. The goal of the total system is to automatically index and retrieve spoken documents. This is consistent with the goal of the retrieval component. The retrieval model makes use of probabilistic quantities that need to be estimated from the spoken documents. This is what the modified speech recognizer now does. Thus the goal of the recognizer is now consistent with the retrieval component and with the goal of the overall system.

The paper is organized as follows. First, we briefly describe our novel probabilistic retrieval model. Next, we describe how the indexing terms are extracted from the speech signal and mention several ways to compute the desired term occurrence probabilities including modifying the speech recognizer objective function. Then, we evaluate the retrieval performance of the integrated approach using a collection of radio broadcast news data. Finally, we summarize the work and close with some conclusions.

\( ^1 \) This research was supported by DARPA under contract N66001-99-1-8904, monitored through NCCOSC.
2. PROBABILISTIC RETRIEVAL MODEL

Our probabilistic information retrieval model scores documents based on the relative change in the document likelihoods, expressed as the likelihood ratio of the conditional probability of the document given the query, \(p(D|Q)\), and the prior probability of the document before the query is specified, \(p(D)\):

\[
S(D, Q) = \frac{p(D|Q)}{p(D)}
\]

The idea is that documents that become more likely after the query is specified are probably more useful to the user and should score better and be ranked ahead of those documents whose likelihoods either stay the same or decrease. We can decompose this likelihood ratio score into more easily estimated components using Bayes’ Rule and rewrite (1) as:

\[
S(D, Q) = \frac{p(Q|D)}{p(D)} \frac{p(D)}{p(Q)} = \frac{p(Q|D)}{p(Q)} \frac{p(D)}{p(Q|D)}
\]

where \(p(Q|D)\) is the probability of query \(Q\) given document \(D\) and \(p(Q)\) is the prior probability of query \(Q\). These probabilities are computed using statistical language modeling techniques. We assume that the query is drawn from a multinomial distribution over the set of possible terms in the corpus so that \(p(Q|D)\) and \(p(Q)\) can be modeled as a product of term probabilities, \(p(t|D)\) and \(p(t)\), over the terms \(t\) in query \(Q\). To address the sparse training data issue, we use Good-Turing methods to estimate \(p(t)\) and a back-off mixture model to estimate \(p(t|D)\) [1]. The resulting retrieval score measure is:

\[
S(D, Q) = \prod_{t \in Q} \left( \frac{\alpha p(t|D)}{p(t)} + \left(1 - \alpha\right) \frac{p(t)}{p(t)} \right)^{q(t)}
\]

where \(q(t)\) is the number of times term \(t\) occurs in query \(Q\), and \(\alpha\) is the back-off mixture weight which is estimated dynamically for each query and automatically using the EM algorithm to maximize the likelihood of the query.

Automatic relevance feedback is a well-established method for improving retrieval performance [6]. It works by running a second retrieval pass using a query constructed by modifying the original query using information from the top scoring documents obtained from a preliminary retrieval pass. We extend our basic retrieval model to include an automatic feedback processing stage by developing a new query reformulation algorithm that is specific to our probabilistic model. The algorithm’s objective is to increase the likelihood ratio score of a joint document composed of the top-ranked documents from the preliminary retrieval pass. This is done by removing certain terms from the original query and adding new terms from the top-ranked documents with appropriate term weights. Hopefully, improving these scores will lead to improved retrieval performance. A complete description of our probabilistic retrieval model can be found in [3].

3. EXTRACTING INDEXING TERMS FROM SPEECH
3.1. Subword Unit Representations

We use subword unit representations as an alternative to word units generated by either keyword spotting or continuous speech recognition. The use of subword units in the recognizer constrains the size of the vocabulary needed to cover the language; and the use of subword units as indexing terms allows for the detection of new user-specified query terms during retrieval [4]. Subword units consisting of overlapping, fixed-length, phone sequences ranging from \(n=2\) to \(n=6\) in length with a phone inventory of 41 classes are used. These subword units are derived by successively concatenating the appropriate number of phones from phonetic transcriptions of the spoken documents. Examples of \(n=1\) and 3 phoneme subword units are shown in Table 1.

<table>
<thead>
<tr>
<th>Subword Unit</th>
<th>Indexing Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>word</td>
<td>weather forecast</td>
</tr>
<tr>
<td>1phon (n=1)</td>
<td>wh d n er f ow r k ae s t</td>
</tr>
<tr>
<td>3phon (n=3)</td>
<td>ow wh d n er f ow r k ae s t</td>
</tr>
</tbody>
</table>

Table 1: Examples of nphone subword unit indexing terms.

3.2. Phonetic Speech Recognizer

The phonetic transcriptions of the spoken documents are generated by a phonetic speech recognizer trained and tuned to operate on the radio broadcast news domain [5]. The recognizer is based on the probabilistic segment-based SUMMIT speech recognition system. It uses context-independent segment and context-dependent boundary (segment transition) acoustic models. Acoustic feature vectors consisting of Mel-frequency cepstral coefficients (MFCCs), difference cepstra, energy, and duration are derived from the speech signal and used in the acoustic models. The distribution of the acoustic features are modeled using Gaussian mixtures. A two pass search strategy is used during recognition: a forward Viterbi search using a statistical bigram language model followed by a backwards A* search using a higher order statistical n-gram language model. The system has a phonetic recognition error rate of 35.0% on an independent development data set.

3.3. Computing Term Occurrence Probabilities

We now describe several methods for estimating \(p(t|D)\), the probability that term \(t\) occurs in spoken document \(D\). This includes using the top one recognition hypothesis, using the \(N\)-best recognition hypotheses, using an expanded term set approach, and modifying the recognizer to compute the term occurrence probabilities directly. These term occurrence probabilities will be used directly in the probabilistic retrieval model.

The simplest approach is to just use the top one recognition hypothesis. In this case, the phonetic recognizer outputs the most likely phone sequence for each document. The appropriate phonetic subword unit indexing terms are generated from the phonetic transcription. And the term counts in each document are used to estimate the term occurrence probabilities:

\[
p_1(t|D) = c_1(t) / \sum c_1(\tau)
\]

where \(c_1(\tau)\) is the number of times term \(\tau\) occurs in document \(D\).

A potentially better estimate of the term occurrence probabilities may be obtained by including additional recognition hypotheses. This can be done by using the \(N\)-best recognition hypotheses, instead of just the top one hypothesis. In this case, the phonetic recognizer outputs the top \(N=100\) phone sequences for each document. For each of the \(N\) hypothesized phonetic transcriptions, the appropriate phonetic subword unit indexing terms are generated. The term counts in this “expanded” document are then used to estimate the term occurrence probabilities:

\[
p_2(t|D) = \sum_{\tau=1}^{N} c_2^\tau(t) / \sum_{\tau=1}^{N} \sum_{\sigma} c_2^\sigma(\tau)
\]

where \(c_2^\tau(t)\) is the number of times term \(t\) occurs in the \(n^\text{th}\) transcription for document \(D\). This estimate reflects the belief that if a term appears in many of the top \(N\) hypotheses, it is more likely to have actually occurred than if it appears in only a few.
Another way to estimate the term occurrence probability is
to incorporate near-miss or approximate match terms. This
can be done by first expanding the term t to include a larger set of
possible realizations of the term \( t^* \) and then summing, over
all members of this expanded set, the probability that t can be
realized as \( t^* \). The occurrence probability of term t in document
\( D_k \), \( p(t|D_k) \), can therefore be computed according to:

\[
p_{\delta}(t|D_k) = \sum_{t^*} p(t|t^*|D_k) = \sum_{t^*} p(t|t^*|D_k) p(t^*|D_k)
\]

where \( p(t|t^*|D_k) \) is an appropriate measure of the probability
that t can be realized as \( t^* \) in document \( D_k \), and the summation
is over all possible realizations, \( t^* \), of term t. The occurrence
probability of term \( t^* \), \( p(t^*|D_k) \), can be estimated using either
Equations 4 or 5 described above. In addition to errors in the doc-
ument transcriptions, this approach also allows for the handling of
synonyms (in a word based system, for example) in a principled
way by summing over an appropriately expanded set of possible
equivalents, \( t^* \), for each original term, t. Information about
the error characteristics of the speech recognizer (i.e., the recognition
error confusion matrix) is used in estimating \( p(t|t^*|D_k) \).

We note that this "expanded term set" approach is very similar to
the approximate matching procedure described in [2].

Finally, we can modify the speech recognizer so that it can output
the \( p_{\delta}(t|D_k) \) probabilities directly. This can be accom-
plished by changing the recognizer’s objective function. Instead
of searching for the most likely phonetic sequence, we want
the recognizer to compute estimates of the probability that indexing
term \( t \) occurred in the given speech message \( D_k \), i.e., \( p(t|D_k) \).
Ideally, we would like to compute this quantity by considering
all possible phonetic sequences \( W \) of the document \( D_k \), finding
all occurrences of the term \( t \) in each sequence \( W \), determining
the probability of each of these occurrences of \( t \), summing those
probabilities to get the expected number of times \( t \) occurred
in the document, and then normalizing it by the total number of
term occurrences in the document to obtain \( p(t|D_k) \).

However, because we cannot consider the exponential num-
ber of possible sequences \( W \), we need to make some approxima-
tions. First, we limit the number of possible phone sequences by
considering only those retained in the phone lattice created by the
Viterbi recognition search. An example phone lattice is shown in
Figure 1. The lattice is a connected acyclic graph where each
node corresponds to a phone hypothesis and has associated with
it a phone label \( p \) for the segment \( s \), the start time \( b_s \) of the
segment, the end time \( e_s \) of the segment, a score \( \delta_s(p) \) repre-
senting the likelihood of the best path from the beginning of the
utterance to the current phone (node), and links (arcs) to possible
following phones (nodes). The \( \delta \) score is computed by the Viterbi
search algorithm. The x-axis represents time and is marked with
possible segment boundary times \( (b_1, b_2, b_3, \ldots) \). Boundaries are
locations in time where the phonetic segments \( s \) are allowed to
start and end. A second approximation is that instead of consider-
ing the term occurrences on all phone sequences represented in
the lattice (which is still a very large number), we only consider
term occurrences on the locally most likely phone sequences. For
each possible segment boundary, we consider all possible phones
that can terminate at that boundary. And for each phone, we find
the term along the most likely phone sequence that terminates at
that phone. A third approximation is in the computation of the
probability of the term occurrence. We take the likelihood score
of the phone sequence corresponding to the term occurrence and
then normalize it to estimate the term occurrence probability.

To generate the phonetic subword unit indexing terms \( t \) and to
estimate the associated occurrence probabilities \( p(t|D_k) \) we use
the following procedure (illustrated in Figure 1.) The first step
is to generate the set of possible indexing terms. This is done
by running local backtraces from all possible phonetic segments
from all possible boundaries. For example, a backtrack of length
\( n=3 \) starting at boundary \( b_5 \) and segment \( s_5 \) with phone label
\( p_0 \) results in the following phone sequence: \( [p_2, p_3, p_5, p_0] \) and
the following tuple: \( \{ p_2, p_3, p_0, b_5, (\delta_{b_5}(p_0)/\delta_{b_5}(p_2)) \} \).
The \( \delta_{b_5}(p_0) \) score corresponds to the likelihood of the best path from
the beginning of the utterance to segment \( s_5 \) ending at bound-
ary \( b_5 \) with phone label \( p_0 \). This best path is equivalent to the
best path from the beginning of the utterance to segment \( s_2 \) end-
ing at boundary \( b_2 \) with phone label \( p_2 \). Plus the following path:
\( (s_3, p_3) \rightarrow (s_4, p_2) \). To determine a score cor-
responding just to the indexing term \( t \), i.e., the phone
sequence \( p_3 \), \( p_0 \), we can take the ratio of the two scores:
\( \delta_{b_5}(p_0)/\delta_{b_5}(p_2) \). In the second step, the term scores are appro-
priately normalized to estimate the term occurrence probability:
\( p(t|D_k) \). First, the scores are normalized over ending boundaries
so that the scores of all terms that end at a specified boundary \( b_k \)
sum to one. Next, the scores are normalized over all the terms that
occur in the document to obtain the final occurrence probability
\( p_{\delta}(t|D_k) \) for term \( t \).

4. RETRIEVAL EXPERIMENTS

4.1. Speech Data Corpus

The speech data used in this work consists of FM radio broad-
casts of the NPR “Morning Edition” news show [4]. The data is
recorded off the air, orthographically transcribed, and partitioned
into separate news stories. The data is broken up into two sets, one
for training and tuning the speech recognizer and another for use
as the spoken document collection for the information retrieval
experiments. The speech recognition training set consists of 2.5
hours of clean speech from 5 shows and the development set con-
ists of one hour of data from one show. The spoken document
collection consists of 12 hours of speech from 16 shows partti-
tioned into 384 separate news stories. Each story averages 2 min-
utes in duration and typically contains speech from multiple noise
conditions. A set of 50 natural language text queries and associ-
ated relevance judgments on the message collection are created
to support the retrieval experiments. The queries are created from
the story “headlines” and are relatively short, each averaging 4.5
words. Each query has an average of 6.2 relevant documents.
4.2. Information Retrieval Performance

We now evaluate the performance of the integrated spoken document retrieval approach. We examine the four different methods for estimating the term occurrence probabilities, $p(t|D_k)$, described in Section 3.3. Figure 2 plots the retrieval performance measured in mean average precision (mAP) for different length ($n = 2, \ldots, 6$) phonetic subword units.

First, the performance of the baseline system implemented using the standard vector space retrieval model is plotted (base). Next, the performance of the probabilistic retrieval model (with automatic feedback) using the top one recognition hypothesis to estimate the term occurrence probabilities $p(t|D_k)$ (4) is plotted (top1). We note that the performance of the probabilistic retrieval model is significantly better than the baseline retrieval model.

Performance using the $N=100$ N-best recognition hypotheses to estimate $p(t|D_k)$ (5) is plotted next (nbest). Performance is slightly but consistently improved over that of using just the top one recognition hypothesis. The use of alternative recognition hypotheses allows additional terms to be included in the document representation and increases the chance of capturing the correct terms. The use of multiple hypotheses also permits a better estimate of the occurrence probability of the hypothesized terms: the more often a term appears in the top $N$ hypotheses, the more likely it is to have actually occurred.

Next, performance using the expanded term set approach to compute term occurrence probabilities $p(t|\mathbf{D})$ (6) is shown (expand). Performance of the short subword unit ($n=2$) gets worse. This is due to an increase in the number of spurious matches caused by the expanded set of terms. The additional terms are likely to match terms that occur in many of the documents due to the short length of the units and the small number of possible terms ($4^2 = 1681$). The performance of the longer subword units ($n=3,4,5,6$), however, are significantly improved. In this case, the expanded set of terms are allowing matches between the clean query terms and the noisy document terms but the longer subword unit sequence length makes it more difficult to get spurious matches. We note that the length $n=3$ subword units now outperform the length $n=3$ units. Also, the performance of the longer subword units ($n=5,6$) are now much closer to the medium length units ($n=3,4$) than before. Previously performance dropped off rapidly as the subword units got longer (e.g., the baseline or top1 curves). This is no longer the case. With the expanded set of terms, the issue of longer subword units being too specific and not matching enough terms has become less of an issue.

Finally, the use of term occurrence probabilities, $p(t|\mathbf{D})$, computed directly by the recognizer is shown (termprob). Performance of the short subword unit ($n=2$) is poor and is actually worse than using expanded term sets. The problem of spurious matches is magnified in this case because even more term possibilities are created when using the term probability approach. The performance of the other subword units ($n=3,4,5,6$), however, are all improved and are better than using the expanded term set approach. Similar to the behavior we saw above with expanded term sets, the additional document terms generated by the term probability approach are allowing more matches with the clean query terms but the longer subword unit sequence length constrains the number of spurious matches, resulting in a net positive effect. As a reference, retrieval performance using subword units derived from error-free text transcriptions is also plotted (text).

Overall, we see that spoken document retrieval performance improves as more sophisticated estimates of the term occurrence probabilities are used. The combined factors of more term hypotheses and improved probability of occurrence estimates to appropriately weigh the additional terms lead to better retrieval performance. The best performance is obtained using term occurrence probabilities computed directly from the speech recognizer: $p(t|\mathbf{D})$. The integrated approach improves spoken document retrieval performance using subword units by 28% over the baseline system: from mAP=0.52 to mAP=0.67.

5. SUMMARY

In this paper, we presented a novel approach to spoken document retrieval where the speech recognition and information retrieval components are more tightly integrated. We use a new probabilistic retrieval model which makes direct use of term occurrence probabilities that can be computed by the recognizer. We described several ways to compute the desired term probabilities including using the top one recognition hypothesis, using $N$-best recognition hypotheses, expanding the term set to include approximate match terms, and modifying the speech recognizer to enable it to output estimates of the term occurrence probabilities directly. We evaluated the performance of the integrated approach using a corpus of radio news data and found that retrieval performance improves as more sophisticated estimates of the term occurrence probabilities are used. The integrated approach improved retrieval performance by 28% over the baseline system.

6. REFERENCES

[3] K. Ng, “A maximum likelihood ratio information retrieval model,” in Eighth Text RETrieval Conference (TREC-8), Gaithersburg, MD, USA, 1999. NIST-SP.