Abstract

In this study, we present the SpeechFind system, an experimental on-line spoken document retrieval system for historical audio archives. As part of an on-going U.S. NSF Digital Library Initiative project, entitled the National Gallery of the Spoken Word (NGSW), SpeechFind is intended to serve as an audio index and search engine for spoken word collections spanning the 20th century with as much as 60,000 hours of audio archives. In this paper, we describe the system architecture of SpeechFind, with focus on audio data transcription and information retrieval components. Using a sample test audio data collection from the past 60 years, an evaluation of individual system components and overall performance is presented.

1. Introduction

During the past decade, there has been impressive advances in both automatic speech recognition and text-based information retrieval. Recently, the development of automatic methods for Spoken Document Retrieval (SDR), continues to emerge as an important research area for both the speech and information retrieval communities. Informally, SDR is the task of retrieving audio information from a large collection of spoken documents to meet a user’s information needs in response to a natural language text query. Successful SDR requires the integration of state-of-the-art automatic speech recognition, information retrieval, and natural language processing technologies. From 1997 to 2000, the annual Text Retrieval Conference (TREC) had included a SDR track with a number of research sites participating. An important observation from the TREC-SDR participants, as well as other researchers, is that retrieval performance degrades only slightly based on automatic transcriptions with Word Error Rates (WER) between 35% and 40%, when compared with retrieval using human transcripts [3]. There are a number of working SDR systems available. Typically, most index and retrieve audio clips using a contemporary broadcast news domain (for example, SpeechBot[4]).

The goal of our on-going five year project entitled the National Gallery of the Spoken Word (NGSW), is to create a significant, fully searchable online database of spoken word collections spanning the 20th century - the first large-scale repository of its kind [11]. NGSW is supported by NSF and is a partnership between a number of research institutions, with Michigan State University (MSU) as the primary institution. The Center for Spoken Language Research (CSLR) at University of Colorado is responsible for providing the recognition and search engine for the 60,000 hours of historical recordings from the last century [2]. Researchers at MSU have been focused on issues relating to digital watermarking [1], standards for digitizing and categorizing NGSW, copyright, distribution, and educational program development. As a digital voice library, NGSW is most recognized for its large collection of audio recordings which spans a wide variety of recording technologies, acoustic environments, speaking styles, epoch specified events, names and places, foreign accents and languages, and evolving grammar and word usage. This wide range of variations present new challenges to current speech and language technologies.

This paper introduces an experimental on-line spoken document retrieval system developed for the NGSW project: SpeechFind [12]. SpeechFind is designed to help world-wide users locate and retrieve audio clips of interest from a large collection of historical recordings for the purposes of research and education. The main components of this system include an audio spider and transcoder, a large vocabulary spoken document transcriber and a web-based public accessible search engine. This paper describes the spoken document transcriber and information retrieval engine for the SpeechFind system, with focus on performance analysis of both speech recognition and information retrieval using a 6-decade sample test data collection. The use of parallel historical documents to improve final retrieval performance is also explored.

This paper is organized as follows: Sec. 2 presents an overview of the SpeechFind spoken document retrieval system; Sec. 3 discusses the transcription of audio archives including both segmentation and recognition; Sec. 4 describes the information retrieval subtask and an overall performance evaluation; Sec. 5 presents a discussion of the present study and directions for future work. Sec. 6 summarizes the paper.

2. SpeechFind System Overview

In this section, we present an overview of the SpeechFind system and briefly describe several key modules. The overall system architecture of SpeechFind is summarized in Fig. 1. The system includes the following modules: an audio spider and transcoder, spoken documents transcriber, “rich” transcription database, and an on-line public accessible search engine.

As shown in the figure, the audio spider and transcoder are responsible for automatically fetching available audio archives from a range of available servers and transcoding the heterogeneous incoming audio files into uniform 16KHz, 16bit linear PCM raw audio data. In addition, for those audio documents with metadata labels, this module also parses the metadata and extracts relevant information into a “rich” transcript database for guiding the future information retrieval.

The SpeechFind transcriber includes two components,
namely the audio segmenter and transcriber. The audio segmenter partitions audio data into manageable small segments by detecting speaker, channel and environmental change points. The transcriber decodes every speech segment into text. If human transcripts are available for any of the audio documents, the segmenter is still applied to detect speaker, channel and environmental changes in a guided manner, with the decoder being reduced to a forced aligner for each speech segment to tag timing information for spoken words. Further details regarding this module are discussed in Sec. 3.

The on-line search engine is responsible for all information retrieval related tasks including a web-based user interface as the front-end, and search and index engines at the back-end. As the audio spider and transcoder, the indexer runs periodically and is activated in an event-driven manner (i.e., indexing the current database when new transcripts or metadata are available). The web-based search engine responds to a user query by launching back-end retrieval commands, formatting the output with relevant transcribed documents that are ranked by relevance scores and associated with timing information, and provides the user with web based page links to access the corresponding audio clips. It should be noted that the local system does not store the entire collection of audio archives, due to both copyright and disk space issues. Instead, SpeechFind only fetchs related audio clips based upon a user's request. Further discussion regarding information retrieval is presented in Sec. 4.

With an enormous effort by researchers from MSU, hundreds of hours of audio archives have already been digitized, and are presently being processed and made accessible by SpeechFind.

3. Transcribing Audio Archives

3.1. Spoken Archives Segmentation

Audio archive segmentation obtains manageable audio blocks for subsequent speech decoding, as well as allows for analysis of the location of speaker(s), channel and environmental change points as useful information to help track audio segments of interest. In addition, further processing can also be incorporated to discard non-speech segments and to cluster homogeneous segments for improved modelling through adaptation.

The segmentation scheme employed here is an iterative $T^2$—Statistic based Bayesian information criterion proposed in an earlier study [10] (We refer this scheme as "$T^2$-BIC" for the remainder of this paper). The feature set used is a 26-dimensional vector (i.e., the 12 static MFCCs and energy plus the first deltas). Initially, the $\lambda$ in Eq. (2) of [10] is set to 1.5 to exclude pseudo segmentations when segmenting the NGSW audio files. If any large segments remain, the value of $\lambda$ is decreased by 0.1 and another round of $T^2$-BIC is applied to the specific large segment. This process is repeated until no segments are longer than 35 seconds. A simple energy-based silence detector is then employed in a guided manner to locate possible silence frames near the $T^2$-BIC break points. These silence frames are picked as the final segmentation points. This silence location processing helps reduce breaks within sentences or phrases. The segmentation step takes about 0.1 times real time for the 6-decade sample test data.

3.2. Spoken Archives Transcription

For SpeechFind, all speech segments are decoded with a large vocabulary recognizer. We are currently using CMU Sphinx3 for this task, but have plans to move to the CSLR Sonic recognizer [5] in the future. The acoustic models contain 5270 GMMs, each of which has 32 mixture Gaussians. Acoustic models are built using a subset of the 200 hours of Broadcast News released by the LDC during 1997 and 1998. The language model is composed of 64K unigrams, 4.7M bigrams, and 15M trigrams. The average decoding speed is about 6.2 times real time on a P4-1.7GHz Linux machine. In establishing the baseline experiments, no model adaptation schemes were applied at this stage, and the first pass decoding output is used as the automatic transcriptions, though a second pass re-scoring using a more complex language model might produce better results.

3.3. Evaluation

To initially evaluate recognition performance, 3.8 hours of sample audio data from the past 6 decades in NGSW is used as the test data. Table 1 summarizes the audio statistics along with the WER averaged for each decade. It is interesting to note that the average WER does not tend to increase as we move back in time, though the Out-Of-Vocabulary (OOV) rate does reflect that trend. Instead, the first 3 decades achieve better recognition accuracy, and the lowest WER is observed for corpora from the 1970’s. This can be partially attributed to the lower average SNR for the recordings used from the 1980s and 1990s. For example, three long audio recordings of 1990’s that contain 2681 words (45.2% of the content from that decade) have an average SNR near 12dB, which produce WERs above 75%, while other recordings with a higher average SNR of 21dB achieve WERs less than 25%. The average SNR of recordings from the 2000s is relatively high, while the audio files are from news conferences regarding the hand counting of votes for the U.S. President in Florida. A considerable portion of the speech is from the audience and is indistinguishable due to the far distance of the speaker source from the recording microphone. As a result, this portion becomes transcriptioned primarily as noise by the recognizer, and as much as 35% of the overall WER is from deletions. While a large portion of the 2000’s WER is from far field audio recordings with audience interference, there is also a significant portion that contains overlapping or competing speaker conversations (i.e., political debates, or news interviews where two or more talkers were attempting to speak simultaneously). It is clear from the wide range of speech recognition performance over the past 60 years from Table 1, that all possible methods for achieving robust

---

The authors thank Rita Singh at CMU and Bhiksha Raj at MERL for their help with Sphinx3.
Table 1: Description and evaluation of sample test audio data.

<table>
<thead>
<tr>
<th>Decades</th>
<th># of Doc</th>
<th>Length of Audio (Min)</th>
<th># of Words</th>
<th>OOV (%)</th>
<th>Avg. SNR(dB)</th>
<th>Avg. WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1950</td>
<td>4</td>
<td>52</td>
<td>6241</td>
<td>1.42</td>
<td>26.63</td>
<td>38.6</td>
</tr>
<tr>
<td>1960</td>
<td>2</td>
<td>17</td>
<td>2142</td>
<td>1.52</td>
<td>21.34</td>
<td>36.7</td>
</tr>
<tr>
<td>1970</td>
<td>2</td>
<td>35</td>
<td>4434</td>
<td>0.81</td>
<td>20.87</td>
<td>25.6</td>
</tr>
<tr>
<td>1980</td>
<td>3</td>
<td>27</td>
<td>3330</td>
<td>0.63</td>
<td>17.97</td>
<td>60.1</td>
</tr>
<tr>
<td>1990</td>
<td>4</td>
<td>47</td>
<td>5951</td>
<td>1.28</td>
<td>14.79</td>
<td>48.0</td>
</tr>
<tr>
<td>2000</td>
<td>3</td>
<td>50</td>
<td>7530</td>
<td>0.78</td>
<td>26.81</td>
<td>59.1</td>
</tr>
</tbody>
</table>

An inherent difference exists between transcribed spoken documents and typical text documents. Automatic transcriptions essentially decode acoustic recordings using the most probable in-vocabulary word sequences. On the other hand, text documents and queries written by humans tend to use a simplified notation. For example, “1960” could be widely used in human-written documents to indicate the year 1960, but it is usually not included in either the dictionary or language models in most state-of-the-art speech recognizers. Hence the audio phrase will appear as “nineteen sixty” in automatic spoken document transcripts.

4. IR Over Automatic Transcripts

The current retrieval engine used in the SpeechFind is a modified version of the MG [9] retrieval system. In this version, the tfidf weighting scheme is replaced with Okapi [7] weighting, and several query and document expansion technologies are incorporated. To ensure sufficient documents from the perspective of IR, the transcript from each recognition segment is treated as a single document. In our case, many historical spoken documents are typically longer than 30 minutes, so the use of small segments as a search unit allows for a more specific user search. In addition, the SpeechFind web interface provides the user access to the detected speech segments and automatic transcripts, and allows the user to preview and listen to any other portions, or the entire audio file that contains the original detected segments.

Table 2 describes the spoken document and query sets used in our evaluation. The 25 test queries were designed by an independent human researcher, based on human transcripts of this test collection, and human relevance assessments were made based on the audio content of the corresponding segments. For indexing, stemming and case folding are performed but no stop words are removed. As our document lengths are considerably shorter than corpora used in the IR literature, we must tune the Okapi parameters for our task. The baseline average precision after spoken transcripts and query normalization is 42.17%; the best performance achieved with $k1 = 0.3$ and $b = 0.75$ for the Okapi weighting scheme. In the following subsections, we report our retrieval performance with experiments on the transcribed spoken documents.

4.1. Spoken Transcripts & Query Normalization

An inherent difference exists between transcribed spoken documents and typical text documents. Automatic transcriptions essentially decode acoustic recordings using the most probable in-vocabulary word sequences. On the other hand, text documents and queries written by humans tend to use a simplified notation. For example, “1960” could be widely used in human-written documents to indicate the year 1960, but it is usually not included in either the dictionary or language models in most state-of-the-art speech recognizers. Hence the audio phrase will appear as “nineteen sixty” in automatic spoken document transcripts. To address this issue, the spoken transcripts and queries are normalized in the SpeechFind system to bridge this gap. Through a predefined dictionary of mappings between “spoken words” and “simplified human notations” the automatic transcripts are filtered, which, for example, replace “N. B. C.” with “NBC”. Using an inverse of a similar dictionary, the queries are filtered as well (e.g., we change the query word “1st” to “first”).

4.2. Query Expansion Using BRF

Query Expansion (QE) is an application that could be used to address the problem of missing query terms directly, or missing term relations indirectly [6]. We first experiment with query expansion using Blind Relevance Feedback (BRF) on the test collection. Here, we consider explicitly adding new terms to the existing query $Q$. We suppose that the top $R$ returned documents are related to the original query in the first round of retrieval, then $T$ expansion terms are chosen according to their Offer Weight (OW) [6] ranking in these $R$ documents. It should be noted that stop-words, defined in a list of 371 common words, that appear in the $R$ documents are first excluded as expansion terms. We experiment with several pairs of $R$ and $T$ and the results are summarized in Table 3. The best result achieved is 44.72% when setting $R = 10$ and $T = 5$.

Table 3: Average precision for query and document expansion.

<table>
<thead>
<tr>
<th>Doc Expan</th>
<th>Query Expan</th>
<th>Avg. Precision(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>$T$</td>
<td>$R$</td>
</tr>
<tr>
<td>2</td>
<td>10%</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>15%</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>10%</td>
<td>–</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>10%</td>
<td>10</td>
</tr>
<tr>
<td>Baseline</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.3. Document Expansion Using PBRF

The main idea behind Document Expansion (DE) [8] is to, given a document, first identify other parallel documents related to those in hand, and then bring “signal” words from the related documents into the present document. To expand spoken documents, we first run the automatic transcription of the speech document as a query on a parallel collection, and then the query documents are expanded using Parallel Blind Relevance Feedback (PBRF).

The effect of document expansion largely depends on the selection of the parallel text collection, which should be related to the spoken corpus. To construct a parallel text collection for our audio recordings, we fetch and parse related historical documents of the 20th century from the web. We also include available human transcripts of our NGSW audio data from same period. The parallel collection contains about 150K words.

In our experiments, we use the same scheme of BRF to expand automatic transcriptions (i.e., using the original spoken transcriptions as the query to search over the parallel collection; all stop words in the spoken documents are not included in the queries). The top R returned documents are assumed to be relevant, then the T expansion terms are chosen to expand the spoken document according to their ranking in terms of a weight scheme in these R documents. Since the transcribed audio segments have considerable length variations, we make T equal to some percentages of the number of terms in each original automatic audio transcription (which achieves better performance than picking a fixed number of terms for all spoken documents).

To rank the candidate terms, we propose the $rtfrw$ weighting scheme:

$$rtfrw(t_i) = rt(t_i) \cdot rw(t_i) = rt(t_i) \cdot \log \left( \frac{(r + 0.5)(N - n - R + r + 0.5)}{(n - r + 0.5)(R - r + 0.5)} \right)$$

where $rw(t_i)$ is the Relevance Weight defined in [6], $r$ is the number of assumed relevant documents where the term $t_i$ occurs, $R$ is the the number of assumed relevant documents for a query, $n$ is the number of documents in the collection where term $t_i$ occurs, and $N$ is the total number of documents in the collection; $rt(t_i)$ is the term frequency of term $t_i$ in the $R$ assumed relevant documents for a query. Candidate expanding terms are selected based on their ranking of $rtfrw$ weight. Again, the stop words are excluded for expansion. Using the $rtfrw$ weight achieves better results than using OW in our experiments. As shown in Table 3, the best performance for using $rtfrw$ weighting is 47.58% (while the best performance using OW is 43.76%, which is not shown in the table).

After obtaining expanded spoken documents, the original queries can also be expanded based on expanded spoken document collections. The results of DE+QE are shown in Table 3. It is clear that with an appropriate choice of $R$ and $T$, the performance of DE using PBRF, and QE using BRF are additive. The combination of DE and QE achieves an average precision of 50.59%, a relative 20% improvement from the baseline precision of 42.17%.

5. Discussion and Future Work

SpeechFind has been established as an experimental platform to perform the task of SDR from historical archives. In the future, the system will be improved through a number of ways. First, the quality of automatic speech transcripts can be boosted by improving the baseline modeling through a set of model adaptation technologies, especially adapting the acoustic models for varied background noises and speakers, and adjusting the language models when topics and decades change dramatically. Moreover, richer information such as accent, stress, emotion and speaker identification contained in spoken segments could also be extracted and used to guide retrieval tasks. Further progresses in the SDR could benefit from improving IR performance. In our task, reliable document categorization could be achieved with the help of metadata associated with some spoken documents, which narrows a search and hence improves the retrieval precision. In addition, a statistical retrieval framework incorporating the uncertainty of automatic transcripts is another interesting research topic.

6. Conclusions

This paper has presented the SpeechFind system, an experimental on-line spoken document retrieval system for a historical archive with 60,000 hours of audio recordings from the last century. We introduced the system architecture, with focus on audio data transcription and information retrieval components. Using a sample test audio data collection from the past 60 years, an evaluation of system components and performance is presented.

7. Acknowledgements

This work was supported by NSF under Cooperative Agreement No. HS-9817485. The authors thank N. Kyrapat for his help in speech recognition evaluation and relevance judgements, M. White for generating queries and relevance judgements, S. Pradhan for helping with Okapi, and extensive thanks to our project partners of Michigan State University for providing sample audio archives and reference transcriptions.

8. References