The Cambridge University Multimedia Document Retrieval Demo System

A.Tuerk†, S.E. Johnson†, P. Jourlin‡, K. Spärck Jones‡ & P.C. Woodland†

†Cambridge University Engineering Department,
Trumpington Street,
Cambridge, CB2 1PZ, UK.
{at233,sej28,pcw}@eng.cam.ac.uk

‡Cambridge University Computer Laboratory,
Pembroke Street,
Cambridge, CB2 3QG, UK.
{pj207,ksj}@cl.cam.ac.uk

Abstract

The Cambridge University Multimedia Document Retrieval Demo System is a web based application that allows the user to query a database of automatically generated transcripts of radio broadcasts that are available on-line. The paper describes how speech recognition and information retrieval techniques are combined in this system and shows how the user can interact with it.

1 Introduction

To provide content-specific access to the vast amount of text data that are available on the Internet, search engines have been developed that operate on text documents of various formats (e.g. html). Since there is an increasing amount of audio data containing speech on the Internet, a similar device is desirable that operates automatically on audio streams without the need of manual transcription. The Cambridge University Multimedia Document Retrieval (CU-MDR) demo system tries to fill this gap.

2 System description

2.1 Overview

The CU-MDR demo system downloads the audio track of news broadcasts from the Internet once a day and adds them to its archive. The audio, which usually comes in RealAudio format, is first converted into standard uncompressed format from which a transcription is produced using our large vocabulary broadcast news recognition engine. This yields a collection of text and audio documents which can be searched by the user. A request by a user triggers a search on the collection of text documents. The returned documents can then be browsed as both text and audio.

2.2 The Speech Recognition Module

The recogniser is similar to the system running in 10 times real time described in (Odell et al., 1999). The audio stream is first split into homogeneous segments. A two pass recogniser is then used to generate the transcriptions for each segment. The first pass produces a rough transcription and then unsupervised model adaptation is used to generate more accurate models for each cluster of acoustic segments. These adapted models are then used with a 4-gram language model to generate the final output. The system is trained on about 150 hours of acoustic training data and 260 million words of broadcast news and newspaper transcriptions. The system gives a word error rate of 15.9% on the 1998 Hub4 broadcast news evaluation data. On the Internet audio used here the run-time is increased due to reduced audio quality but the general level of transcription accuracy remains high.

2.3 The Information Retrieval Module

The information retrieval engine used in the CU-MDR demo is the benchmark system described in (Johnson et al., 2000). Semantic posets (Jourlin et al., 1999) are not automatically included in this system. Instead they are used to suggest additional words that can be added to the original query. Relevance feedback is also available. This allows the user to mark the documents that contain relevant information and have the system suggest additional query words that distinguish those from the non-relevant documents. When activated both query expansion methods bring up a list from which the user can select the words that he/she believes are most useful in expanding the query.

2.4 Interface and User Interaction

2.4.1 Entering Queries The user gains access to the MDR demo by visiting the MDR demo web page using a conventional (preferably 4th generation) browser. After registering with the system the user can query the audio/text database. This can be done in two ways. Either the user queries the system interactively, submitting a request by typing a query into the search field or he/she can specify a set of queries that represent his/her long-term interests. The retrieval engine is run on these long-term queries only on login. If the retriever finds a document that matches one of these queries and which has been added to the data base since the user's last session, the query is highlighted. This feature effectively allows the user to filter the incoming broadcasts.

2.4.2 Presentation of Search Results Once the retriever has returned the results for a particular search, a list of extracts from the returned text documents is created. Each extract is designed to represent the part of the document that is most relevant to the query. This list can be sorted using different criteria, e.g. highest relevance score first, most recent first. Each extract highlights the query words and also shows how often the query words were found in the whole document. The ratio of the relevance score for a document to the score of the top ranking document for the current search is also displayed. The user can listen to the part of the sound source that corresponds to the extract. The whole automatic transcript can be accessed on a separate web page where the user can listen to selected parts of the transcription by highlighting them.

3 Ongoing Work

At the moment the CU-MDR database consists of pre-segmented NPR broadcasts only. Work is ongoing to extend the data base to British English news. Also a windowing system is being developed that allows automatic content dependent segmentation of news broadcasts.

Acknowledgements

This work is in part funded by an EPSRC grant reference GR/L49611.

References

Johnson, S. E., Jourlin, P., Spärck Jones, K., and Woodland, P. C. (2000). Spoken Document Retrieval for TREC-8 at Cambridge University. To appear. In *Proc. TREC-8*, NIST Gaithersburg, MD.

Jourlin, P., Johnson, S. E., Spärck Jones, K., and Woodland, P. C. (1999). General Query Expansion Techniques for Spoken Document Retrieval. In *Proc. ESCA Workshop on Extracting Information from Spoken Audio*, pages 8–13, Cambridge, England.

Odell, J. J., Woodland, P. C., and Hain, T. (1999). The CUHTK-Entropic 10xRT Broadcast News Transcription System. In *Proc. DARPA Broadcast News Workshop*, pages 271–275, Herndon, VA.