Integration of a Large Text and Audio Corpus Using Speaker Identification

Deb Roy Carl Malamud

MIT Media Laboratory
20 Ames Street, E15-388
Cambridge, MA 02139
dkroy@media.mit.edu carl@media.mit.edu

Abstract

We report on an audio retrieval system which lets Internet users efficiently access a large text and audio corpus containing the transcripts and recordings of the proceedings of the United States House of Representatives. The audio has been temporally aligned to corresponding text transcripts (which are manually generated by the U.S. Government) using an automatic method based on speaker identification. This system is an example of using digital storage and structured media to make a large multimedia archive easily accessible.

Introduction

In the United States, the text of proceedings of the two houses of the Congress has long been published in the Congressional Record. No systematic effort has been made, however, to record audio from the floor of the House and Senate. In 1995, the non-profit Internet Multicasting Service (IMS) began sending out live streaming audio to the Internet and making complete digital audio recordings of the proceedings on computer disks. The challenge was to make this massive amount of recorded audio information available to Internet users in a meaningful way.

After investigating a variety of options, we decided to couple the Congressional Record (the text database) to the audio database. The resulting system allows users to efficiently search, browse, and retrieve audio over the Internet. The basic idea is to allow searches on a text transcript and then locate the audio which corresponds to the text search results. Correspondences between the text and audio are generated by automatically identifying the identity of speakers in a recording and aligning speaker transitions in the audio with corresponding speaker transitions in the associated text transcript.

Recently there have been several efforts to build audio retrieval and indexing systems. The most popular approach has been to index audio based on content words using either large vocabulary speech recognition or keyword spotting (Wilpon at al. 1990, Rose, Chang & Lippman 1991, Wilcox & Bush 1991, Glavitsch & Schauble 1992, Jones et al. 1996, James 1996). Other cues including pitch contour, pause locations and speaker changes have also been used (Chen & Withgott 1992, Arons 1996, Roy 1995). In one system the closed caption text of television news broadcasts was aligned to the audio track based on pause locations enabling users to perform searches on text and then access corresponding audio (Horner 1991).

Media retrieval systems will continue to grow in importance as digital archives such as the following become common:

- Radio and television broadcast archives
- Internet sites with speech and music (already quite common)
- Recordings from various local sources such as lecture halls, and courts
- Data collected on wearable computers which record f irst-person media (video, audio and other information)

Since the majority of such archives will include speech, this is a natural application domain for speech processing. By extracting some structure from audio, an archive which is inaccessible due to it's size and the difficulties of searching unstructured audio can become searchable by content. Applications include multimedia content re-use, audio note taking, and content-searchable multimedia archives.

In this paper we describe a novel method based on speaker identification which was used to align the text and audio recordings of the proceedings of the Congress. We also describe the WWW interface which readers are invited to try at http://town.hall.org/Congress/. The resulting system enables Internet users to quickly locate original congressional proceedings which were previously unavailable in audio form.

The Congressional Databases

The Congressional Record includes edited transcripts of the proceedings, manually generated time stamps, results of any votes, and scheduling information about upcoming sessions. The transcripts are originally created live during the proceedings by a human transcriber. Among other things, two types of information recorded by the transcriber are of particular interest for the automatic text to audio alignment task: each speaker transition is recorded, and time stamps spaced every 10 to 45 minutes are entered during long pauses in the proceedings. One of the significant challenges is that the Congressional Record is not a verbatim record of the proceedings. Members have the opportunity to add new material, abridge their remarks, and otherwise edit the transcripts.

The audio database used for the experiments described in this paper contains 132 hours of proceedings of the House of Representatives recorded from January 20 through February 22, 1995. We also collected the corresponding text transcripts in electronic form.

The Congressional Database Retrieval System

Figure 1 shows the main components of the audio retrieval system. The text and audio databases described in Section 2 are shown at the top of the figure. The World Wide Web (WWW) interface enables users to constrain searches using a variety of parameters (see Section 5 for more details). The search parameters are used to locate selections of text from the text database. The text search engine includes a parser which extracts information about the date, time, and speaker identity from the text databases and uses this information to enforce some of the user specified search constraints. The search engine returns pointers to speaker transition points within the text which indicate search matches. The text to audio alignment system then provides pointers into the audio database which correspond to the selected text. The WWW interface also provides both audio playback and a text display so users can interactively skim both the text and the audio in real time over the Internet (real time audio play back is supported over the Internet multicasting backbone using several popular audio transfer protocols including VAT, Real Audio and Xing).

Text To Audio Alignment

A key component of the audio retrieval system is the text to audio alignment system which performs an automatic time alignment of the audio and text databases. One method of performing the alignment might be to run a large vocabulary speech recognizer on the audio and align the text output of the recognizer to the text transcript. This approach is difficult because the transcriptions often stray significantly from the verbatim words of the audio. Additionally, the original audio recordings have variable signal to noise ratios which makes speech recognition difficult (speakers talk into an open microphone mounted on a floor stand; the microphone occasionally picks up considerable background noise from other people present in the chamber.).

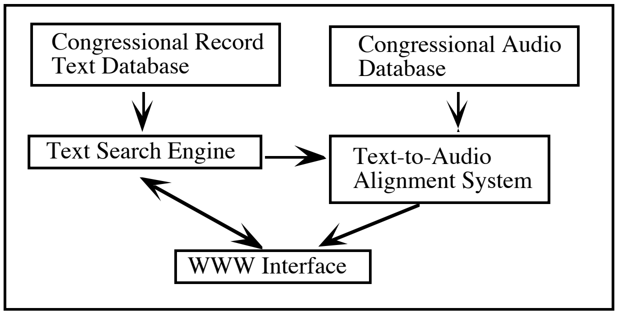

Congressional Record Text Database
Congressional Audio Database
Text Search Engine
Text-to-Audio Alignment System
WWW Interface

Figure 1: The Congressional Audio Retrieval System

Rather than attempt to align text and audio using speech recognition, our approach is to use speaker identification. We extract the sequence of speakers from the text transcript. We then use acoustic models of the speakers to locate points in audio where speaker transitions occur. We can then find correspondences between the text and audio at these points of speaker change. In addition to the speaker sequence, we also use the time stamps to further constrain the speaker identification process.

We have implemented the alignment system shown in Figure 2. The text parser extracts the sequence of speakers and time stamps. Although the Congressional Record was not designed to be machine readable, its structured format allowed us to find the names of the speakers with fairly high accuracy. The time stamps are also well marked and can be extracted from the text easily but were found to be accurate only within a range of about two minutes.

We have built Gaussian models for the voice of each of 435 members of the House of Representatives based on cepstral features. The models are used in conjunction with the speaker sequence and time stamps (extracted from the text transcript) to constrain a Viterbi alignment of the speaker transition points in the text and the audio. For technical details of the alignment process see (Roy & Malamud 1997). The Viterbi alignment results in a coupling of the audio and text at each speaker change boundary.

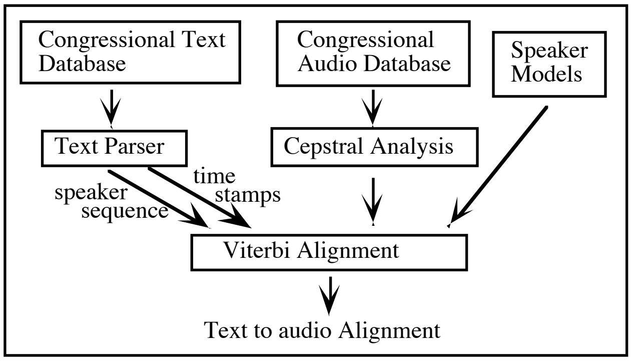
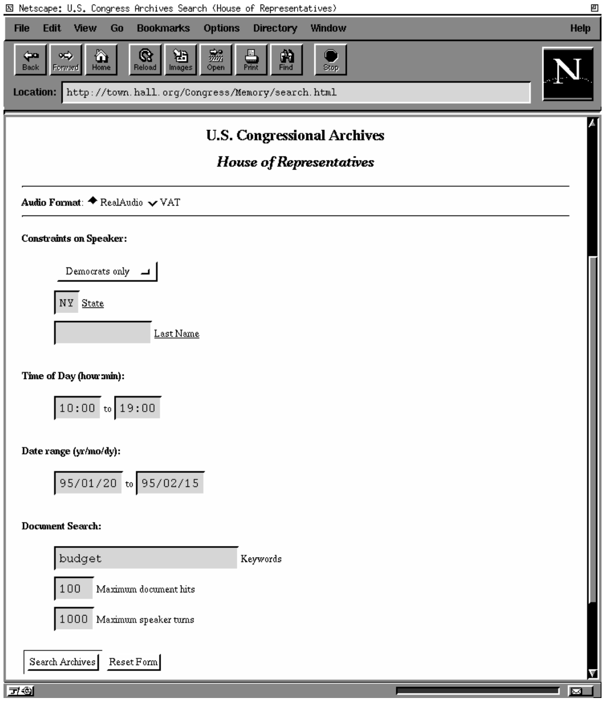


Figure 2: Estimating temporal alignment of text and audio

The Search And Browse Interface

We have built two WWW interfaces for accessing the audio database over the Internet. The primary interface is a search form with which the user can search for audio segments constrained by several criteria including keywords, name of speaker, political party of speaker, speaker's home state, date range, time range (specify range of times within a day). Figure 3 shows a search page in which the user has requested a search for all speeches by New York Democrats who spoke within the date range 95/01/20 to 95/02/15 and whose speech contained the keyword "budget".


Netscape: U.S. Congress Archives Search (House of Representatives)
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
N

Location: http://town.hall.org/Congress/Memory/search.html

U.S. Congressional Archives

House of Representatives

Audio Format: ◆ RealAudio ✓ VAT

Constraints on Speaker:

Democrats only

NY State

Last Name

Time of Day (hour:min):

10:00 to 19:00

Date range (yr/mo/dy):

95/01/20 to 95/02/15

Document Search:

budget Keywords

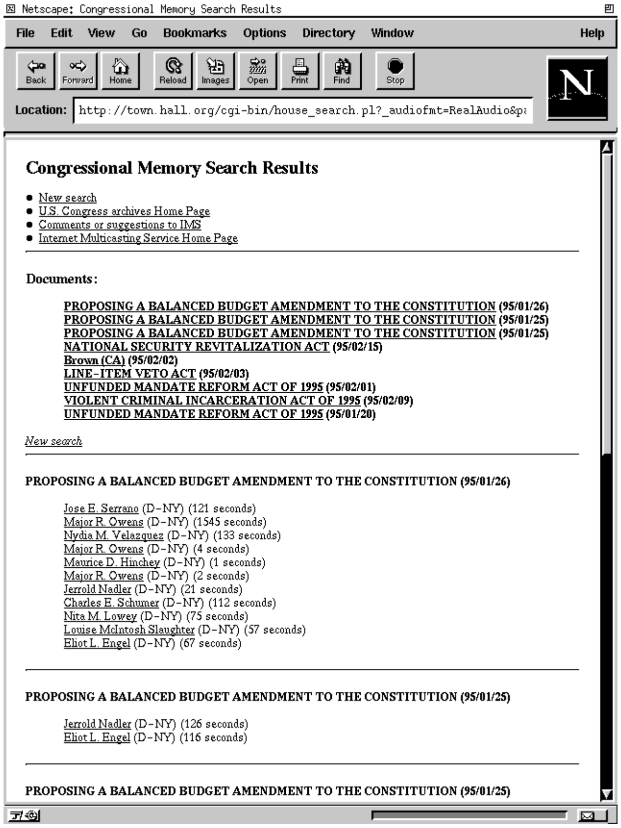
100 Maximum document hits

1000 Maximum speaker turns

Search Archives | Reset Form

Figure 3: WWW Search interface to the Congressional proceedings corpus.

The search engine in the web server finds all speeches in the text corpus which meet the search criteria and presents them as shown in Figure 4. This page first lists all transcript documents which contain hits, and then list each speaker (note that all speakers are New York Democrats as specified in the search). The user may then follow any link from this page to read the full text and listen to the corresponding audio of each speech. Figure 5 shows a typical web page when a link from one of the speakers is followed. The scrollable window at the top contains the transcript of the speech, and the control buttons at the bottom enable interactive playback and navigation of the audio. Assuming the Viterbi alignment was successful for this speaker, simply hitting the play audio will play the audio from the beginning of the speech.


Netscape: Congressional Memory Search Results
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
N

Location: http://town.hall.org/cgi-bin/house_search.pl?_audiofmt=RealAudio&pa

Congressional Memory Search Results

- New search
- U.S. Congress archives Home Page
- Comments or suggestions to IMS
- Internet Multicasting Service Home Page

Documents:

PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/26)
PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/25)
PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/25)
NATIONAL SECURITY REVITALIZATION ACT (95/02/15)
Brown (CA) (95/02/02)
LINE-ITEM VETO ACT (95/02/03)
UNFUNDED MANDATE REFORM ACT OF 1995 (95/02/01)
VIOLENT CRIMINAL INCARCERATION ACT OF 1995 (95/02/09)
UNFUNDED MANDATE REFORM ACT OF 1995 (95/01/20)

New search

PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/26)

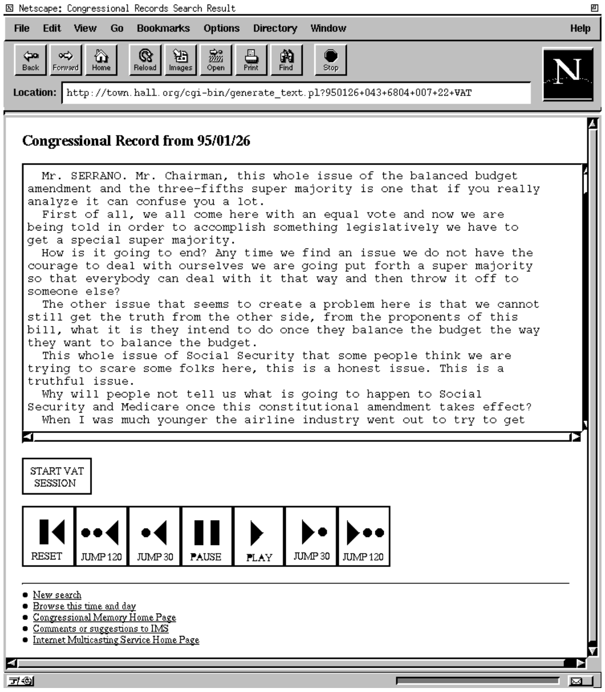
Jose E. Serrano (D-NY) (121 seconds)
Major R. Owens (D-NY) (1545 seconds)
Nydia M. Velazquez (D-NY) (133 seconds)
Major R. Owens (D-NY) (4 seconds)
Maurice D. Hinchey (D-NY) (1 seconds)
Major R. Owens (D-NY) (2 seconds)
Jerrold Nadler (D-NY) (21 seconds)
Charles E. Schumer (D-NY) (112 seconds)
Nita M. Lowey (D-NY) (75 seconds)
Louise McIntosh Slaughter (D-NY) (57 seconds)
Eliot L. Engel (D-NY) (67 seconds)

PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/25)

Jerrold Nadler (D-NY) (126 seconds)
Eliot L. Engel (D-NY) (116 seconds)

PROPOSING A BALANCED BUDGET AMENDMENT TO THE CONSTITUTION (95/01/25)

Figure 4: Example search result lists all speeches which meet the specified search criteria.


Netscape: Congressional Records Search Result
File Edit View Go Bookmarks Options Directory Window Help
Back Forward Home Reload Images Open Print Find Stop
N
START VAT SESSION
RESET
JUMP 120
JUMP 30
PAUSE
PLAY
JUMP 30
JUMP 120

Location: http://town.hall.org/cgi-bin/generate_text.pl?950126+043+6804+007+22+VAT

Congressional Record from 95/01/26

```
  Mr. SERRANO. Mr. Chairman, this whole issue of the balanced budget
amendment and the three-fifths super majority is one that if you really
analyze it can confuse you a lot.
  First of all, we all come here with an equal vote and now we are
being told in order to accomplish something legislatively we have to
get a special super majority.
  How is it going to end? Any time we find an issue we do not have the
courage to deal with ourselves we are going put forth a super majority
so that everybody can deal with it that way and then throw it off to
someone else?
  The other issue that seems to create a problem here is that we cannot
still get the truth from the other side, from the proponents of this
bill, what it is they intend to do once they balance the budget the way
they want to balance the budget.
  This whole issue of Social Security that some people think we are
trying to scare some folks here, this is a honest issue. This is a
truthful issue.
  Why will people not tell us what is going to happen to Social
Security and Medicare once this constitutional amendment takes effect?
  When I was much younger the airline industry went out to try to get
```

- New search
- Browse this time and day
- Congressional Memory Home Page
- Comments or suggestions to IMS
- Internet Multicasting Service Home Page

Figure 5: Display page for a speech; top window contains transcript, buttons at bottom enable interactive playback of corresponding audio.

All audio alignment is precomputed by the text to audio alignment system off-line so the system response time is quick.

In cases there the audio alignment has failed, a secondary browse interface can be brought up by the user which displays a list of temporally close segments before and after the initially selected audio segment. Since errors in the alignment are typically within a few minutes of the actual location (the worst case error is limited to the time stamp interval or that section of the Congressional Record), the user can use the browser to quickly locate the information of interest.

Conclusions And Future Work

We have presented a system for aligning the audio and text of the proceedings of the U.S. House of Representatives. The alignment effectively couples the text and audio databases in a manner which enables efficient search and browsing of hundreds of hours of audio. The resulting audio retrieval system has been deployed experimentally on the Internet since October, 1995.

We plan to add other methods of audio analysis including keyword spotting and prosody analysis to extract further structure from audio recordings. This additional structure can be used to improve navigation within segments of audio located by the speaker identification system. We also plan to use related methods for extracting low level structure from video. For example jump cuts are easily found by automatic systems, and the presence of people (by looking for flesh tones) is also relatively simple and robust.

The first author is currently working on two new applications using the structured media retrieval methods described in this paper and in (Roy 1995). The first is to make the proceedings of the European Union Summit available on the Internet, and the second application is in wearable computing; We are building a wearable computer which continuously records the audio environment of the user. The resulting audio archive will be structured using various methods and made accessible to the user through the wearable, providing a form of augmented memory.

References

Wilpon, J.G., Rabiner, L.R, Lee, C. and Goldman, E.R.. 1990. Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models. IEEE Trans. on Acoustics, Speech, and Signal Processing 38, 11, pp. 1870-1878.

Rose, R.C., Chang, E.I., and Lippman, R.P. 1991. Techniques for Information Retrieval from Voice Messages. In Proc. ICASSP, pp. 317-320, IEEE, Toronto.

Wilcox, L. and Bush, M. 1991. HMM-based Wordspotting for Voice Editing and Indexing. In Eurospeech ‘91, pp. 25-28.

Glavitsch, U. and Schauble, P. 1992. A System for Retrieving Speech Documents. In 15th Annual International SIGIR ‘92, ACM, New York, pp. 168-176.

Jones, G. J. F., Foote, J.T., Sparck Jones, K., and Young, S.J.. 1996. Robust Talker-Independent Audio Document Retrieval. In Proc. ICASSP, pp. 311-314, IEEE, Atlanta.

James, D.A.. 1996. A System for Unrestricted Topic Retrieval from Radio News Broadcasts. In Proc. ICASSP, IEEE, Atlanta.

Chen, F. and Withgott, W. 1992. The Use of Emphasis to Automatically Summarize a Spoken Discourse. In Proc. ICASSP, pp. 229-232, IEEE, San Francisco.

Arons, B. 1994. Speech Skimmer: Interactively Skimming Recorded Speech. Ph.D. thesis, MIT Media Laboratory,.

Roy, D. 1995. NewsComm: A Hand-Held Device for Interactive Access to Structured Audio. Masters thesis, MIT Media Laboratory.

Horner, C. 1991. NewsTime: A Graphical User Interface to Audio News. Masters thesis, MIT Media Laboratory.

Roy, D. and Malamud, C. 1997. Speaker Identification based Text to Audio Alignment for an Audio Retrieval System. To appear in Proc. ICASSP, IEEE, Munich,.