SPEAKER IDENTIFICATION BASED TEXT TO AUDIO ALIGNMENT FOR AN AUDIO RETRIEVAL SYSTEM

Deb Roy[1] *Carl Malamud*[1,2]

[1]MIT Media Laboratory, 20 Ames Street, Cambridge, MA 02139 USA
[2]Internet Multicasting Service, 31 Madison Street, Cambridge, MA 02138 USA
dkroy@media.mit.edu, carl@media.org

ABSTRACT

We report on an audio retrieval system which lets Internet users efficiently access a large audio database containing recordings of the proceedings of the United States House of Representatives. The audio has been temporally aligned to text transcripts of the proceedings (which are manually generated by the U.S. Government) using a novel method based on speaker identification. Speaker sequence and approximate timing information is extracted from the text transcript and used to constrain a Viterbi alignment of speaker models to the observed audio. Speakers are modeled by computing Gaussian statistics of cepstral coefficients extracted from samples of each person's speech. The speaker identification is used to locate speaker transition points in the audio which are then linked to corresponding speaker transitions in the text transcript. The alignment system has been successfully integrated into a World Wide Web based search and browse system as an experimental service on the Internet.

1. 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, literally hundreds of hours, available to Internet users in a meaningful way. After investigating a variety of options, it was decided to couple the Congressional Record (the text database) to the audio database that we accumulated. The resulting system allows users to efficiently search, browse, and retrieve audio over the Internet.

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 [1, 2, 3, 4, 5, 6]. Other cues including pitch contour, pause locations and speaker changes have also been used [7, 8, 9]. 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 [10].

Audio retrieval systems will continue to grow in importance as digital archives become more common.[1] Since the majority of such archives will consist of 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. The resulting system enables Internet users to quickly locate original congressional proceedings which were previously unavailable in audio form.

2. 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 audio was sampled at 16 kHz with a 16 bit digitizer. The recordings consist mostly of speech from the 435 male and female members of the House of Representatives. The recordings also contain occasional background speech, applause, laughter, and other noise.

[1] For example most large radio broadcasters in the U.S. are switching to digital and will begin to accumulate massive digital audio archives.

3. 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).

4. 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[2] which makes speech recognition difficult.

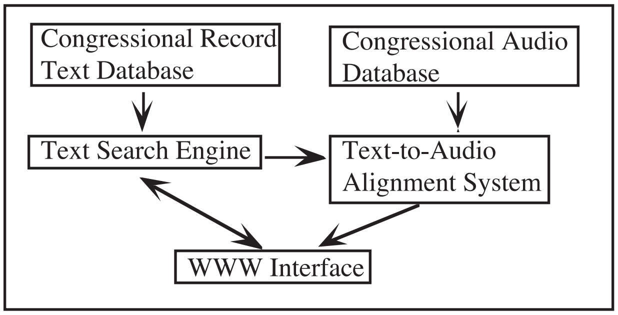
Congressional Record Text Database
Congressional Audio Database
Text Search Engine
Text-to-Audio Alignment System
WWW Interface

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 (described below) 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 as described later in this paper.

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, it uses fairly standard formats for indicating speaker changes. There were errors in parsing this information since the format is not always followed in the transcript; this is discussed in Section 5. The time stamps are also well marked and can be extracted from the text but were found to be accurate only within a range of about two minutes.

We used methods similar to [11] to build acoustic models for each of the 435 members of the House of Representatives. First we manually located two 30 second segments of speech from each person spoken on two different days. For each 60 second sample of audio, we then computed 12th order cepstral coefficients using a 32 ms Hamming window with 16 ms overlap, and then computed the mean vector and covariance matrix, and Σ, of these cepstral vectors.

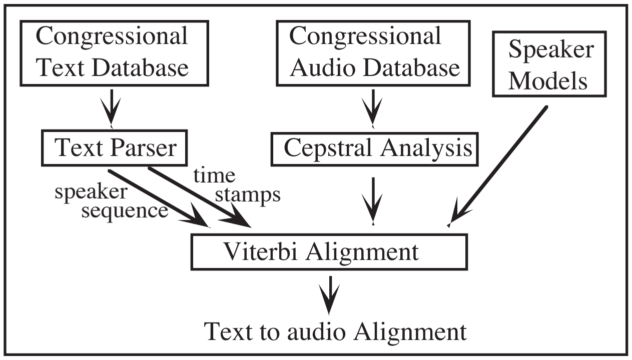

Figure 2: Estimating temporal alignment of text and audio

For a cepstral vector x and Gaussian model with mean and covariance Σ the likelihood that the vector was produced by the model is given by (1):

$$L(x;\ ,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\ )'\Sigma^{-1}(x-\ )\right\} \quad (1)$$

(The vertical bars denote taking the determinant of the enclosed matrix).

Given the audio recording and text record of a day, we obtain three types of information: (i) a matrix of values indicating the likelihood of each frame of audio in the day was generated by each speaker model (by applying Equation (1) to each observed cepstral frame using each speaker model), (ii) the sequence of speakers who spoke in the day (from the text), and (iii) approximate time stamps recorded every 10 to 45 minutes throughout the day (also from the text).

[2] 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.

By constraining the sequence of models which are considered using the speaker sequence extracted from the text, we can solve the alignment problem using the Viterbi algorithm [12]. Specifically, we wish to recover the best speaker sequence Q=(q_1, q_2,...q_t) given the observed sequence of cepstral vectors X=(x_1, x_2,...x_t). The likelihood for a given speaker sequence is:

$$\delta = \prod_{i=1}^{t} L(x_i;\ \ _i \Sigma_i) \qquad (2)$$

Where $_i$ and $\sum_i$ are the mean and covariance computed from the 60 second sample of speaker q_i.

We used the Viterbi algorithm to find the speaker sequence which maximizes δ in Equation (2). We further constrained the Viterbi algorithm with the time stamps from the transcripts which usually occur during long pauses in the audio recordings (this is when the transcriber has time to check the clock and type in the time). To use this information we located pauses in the audio recording by thresholding the short-time energy of the audio using a fixed threshold. We then located the pause in the audio closest to each time stamp and segmented the audio recording into segments of lengths between 10 and 45 minutes. We processed each audio segment separately using the corresponding segment of the speaker sequence list to constrain the Viterbi alignment. This time stamp based segmentation insures a limit on the drift in alignment error since the time stamps act as anchor points; a speaker sequence alignment is forced locally within each segment and the error from one segment cannot effect the error in the next.

We note that since the original time stamps are only accurate to within approximately 2 minutes there is occasionally an error in the assumed speaker sequence at either end of the segments which leads to some alignment errors.

5. RESULTS

In this section we first describe the types of errors we encountered in the system and then present some preliminary speaker alignment accuracy results on a small subset of the corpus.

We have found several causes for errors in the speaker identification process:

- undocumented speaker turns in the text transcripts (for example there are several cases in which the transcript indicates a single speaker transition but in the corresponding audio two speakers argue for control of the floor generating as many as 15 (often overlapping) speaker transitions!)
- non-standard documentation of speaker turns in the transcripts which were not detected by our text parser
- short speaker turns which are overlapped completely by another speaker in the audio recordings
- undocumented speakers for whom we have no voice models (usually clerks who announce administrative details of the proceedings)

Each of these errors may cause the Viterbi alignment to fail. Fortunately the time stamps provide alignment anchors which prevent errors from propagating beyond a single audio segment. In our initial analysis we have found that close to half of the audio segment alignments have errors due to one of these sources. Some directions for future work to alleviate these problems are described at the end of this section.

For this initial analysis, we randomly selected five one-hour audio segments recorded on three different days. Of these segments we found that two of the segments contained one of more of the errors listed above and were not analyzed further. We note that the average error in alignment for all speaker transition points in these two segments was greater than one minute. The three remaining segments of audio contained a total of 49 speaker transitions. Figure 3 summarizes the speaker alignment performance on this set of audio. Any point along the curve indicates the percentage of speaker transitions (vertical axis) which are located with an error less than or equal to a specific alignment error (horizontal axis). For example 57% of the transitions are located by the system with an alignment error of 3 seconds or less; 70% of the transitions are found when a 15 second error is acceptable. We note that for the task at hand (i.e. retrieving specific audio segments from several hundred hours of archives) alignment errors in the tens of seconds still provides a very useful service.

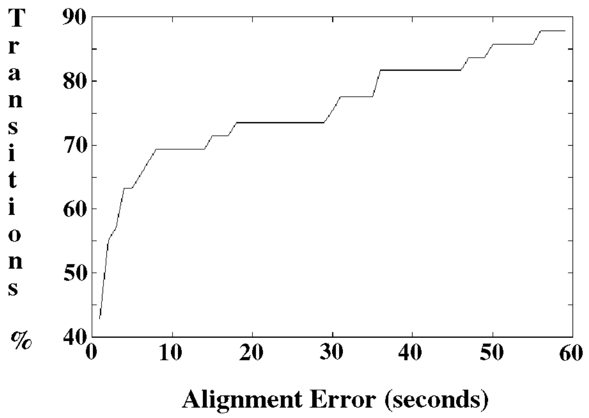

Figure 3: Preliminary results of the speaker alignment system (results shown only for the segments which did not contain errors of the type listed in the text)

This preliminary analysis suggests that we need to soften the constraints derived from the text transcripts so that the speaker sequence produced by the Viterbi alignment does not have to match the text sequence precisely. We also note that the length of text associated with each speaker turn in the transcript can be used as a predictor of the duration of the corresponding audio segment. In future work we plan to derive duration predictions from the text and use them as a further constraint on the alignment procedure.

6. THE SEARCH AND BROWSE INTERFACE

We have build 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).

A full text search is performed on the text database and a list of candidate hits are listed on a web page. The user can click on an entry from the list to go to a display page which has a scrollable window to view the text, and an audio playback interface to listen to the corresponding audio over the network. 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[3], the user can use the browser to quickly locate the information of interest.

7. CONCLUSIONS

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 at the WWW site http://town.hall.org.

8. ACKNOWLEDGMENT

The authors would like to thank Herb Gish for many useful discussions.

9. REFERENCES

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

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

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

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

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

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

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

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

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

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

[11] H. Gish and M. Schmidt, "Text-Independent Speaker Identification." IEEE Signal Processing Magazine, Oct. 1994, pp. 18-32.

[12] L. Rabiner and B. Juang. Fundamentals of Speech Recognition. Prentice Hall, New Jersey, 1993.

[3] The worst case error is limited to the time stamp interval for that section of the Congressional Record.