Audio/Video Navigation with A/V X-Ray

Patrick Nguyen and Milind Mahajan
Speech Research Group
Microsoft Research

Summary

The convergence of Audio/Video (A/V) consumption shifting to the Internet is well underway. Current online technologies fail to realize the full potential of this phenomenon. Unlike text, because A/V may hardly be analyzed at a glance. Discovering new content, gauging search results relevance, or browsing becomes difficult. Our solution, A/V X-Ray, reveals the hidden structure of the A/V content. It performs A/V content analysis which involves: A/V scene analysis and speech recognition combined with story, speaker, topic segmentation and summarization. A fully connected experience is achieved by presenting related web content, which provides context, and collaborative meta-data at a fine granularity.

Motivation

A/V convergence has resulted in an explosion of A/V content available to the user when and where the user wants it. This creates a need for better tools for the user to deal with the amount of available content to locate and consume the content of interest. Keyword based A/V search partially addresses the issues of discoverability of the content. However, A/V search alone is not enough. Evaluation of relevance of the A/V content surfaced through search is non-trivial given the opacity and the linear nature of the A/V content. A/V content is not easily digestible at a glance in a visual snapshot. It is also consumed linearly which makes it time-consuming to sample. Users also tend to interact with the long format A/V content by browsing through the content as opposed to active searching for something specific. Improving this browsing experience is also a significant benefit to the users.

A/V X-Ray

A/V X-Ray is our proposed solution for improving the user experience. **A/V X-Ray reveals the hidden structure of the A/V content** to the user through an effective user interface. It thus enhances the usefulness of A/V search and also complements the A/V search as an alternative way of navigating the A/V content. To achieve this, it performs A/V content analysis which involves: A/V scene analysis and speech recognition combined with speaker segmentation, topic segmentation and summarization. By using all this information, it is immediately apparent from visual inspection of the X-Ray what proportion of time is spent on each topic, which speaker is most loquacious, what the aggregated community reaction and rating is, and what the context is for each video segment.

A/V content is pre-processed by a simple classifier to determine the speech and non-speech regions. Automatic transcript is then provided by processing the speech regions with a large vocabulary ASR. This provides the user a quick way to browse the contents textually to get a rough idea of the content. The transcript is also aligned with the A/V content which provides a way for the user to jump to a specific point in the A/V content.

Automatic topic segmentation creates an organization for the A/V content. A combination of lexical and acoustic features in maximum entropy framework is used to generate the topic segmentation. The level of granularity of the segmentation can be controlled by the user. The organization created by this segmentation can be used for navigation. The user may skip the segments which are not of interest. This can be thought of as intelligent fast-forward. Alternatively, the organization can be used to scan through the program sampling each segment to create a simple A/V summary of the content. Each topic segment is automatically labeled with a few keywords using features such as term frequency and inverse document frequency (tf-idf). This provides the context for the navigation.

Speaker segmentation is performed using bottom-up agglomerative clustering. BIC is used to control the level of clustering. Visual presentation of the speaker turn information provides useful information to the user about the A/V content at a glance. It also provides navigation and context capabilities similar to the topic segmentation.

Integration with meta-data

The history of the web search highlights the importance of the meta-data. Meta-data for the A/V content consists of meta-data provided by the content producer such as: web links, summaries, graphics, and advertisements. This meta-data can be further enhanced by user community by providing commentary, discussions and ratings. X-Ray enriches the A/V user experience by presenting the meta-data to the user in sync with the A/V content. A/V X-Ray encourages the production and consumption of meta-data in an interleaved, interactive fashion. Meta-data, in turn, can enhance the ability to perform automatic content analysis by providing the much needed relevant data. For the users, the ability to add comments is a way to participate and express themselves.

Rich user experience

Meta-data, automatic content analysis and collaborative filtering data generated by community interaction with the content are also used by A/V X-Ray to recommend related A/V content, related information on the web and display related advertisements. Related web information and advertising provides the user with a broader context for the A/V content. Recommendations for the A/V content provide an alternative navigation mechanism. Combining these features with automatic topic segmentation increases the granularity at which they can operate and provide increased benefit for the long-format A/V content.

Illustrations

Included below are images which explain these concepts graphically. The images are only for illustrative purposes and are not meant as the screenshots of the working demo.

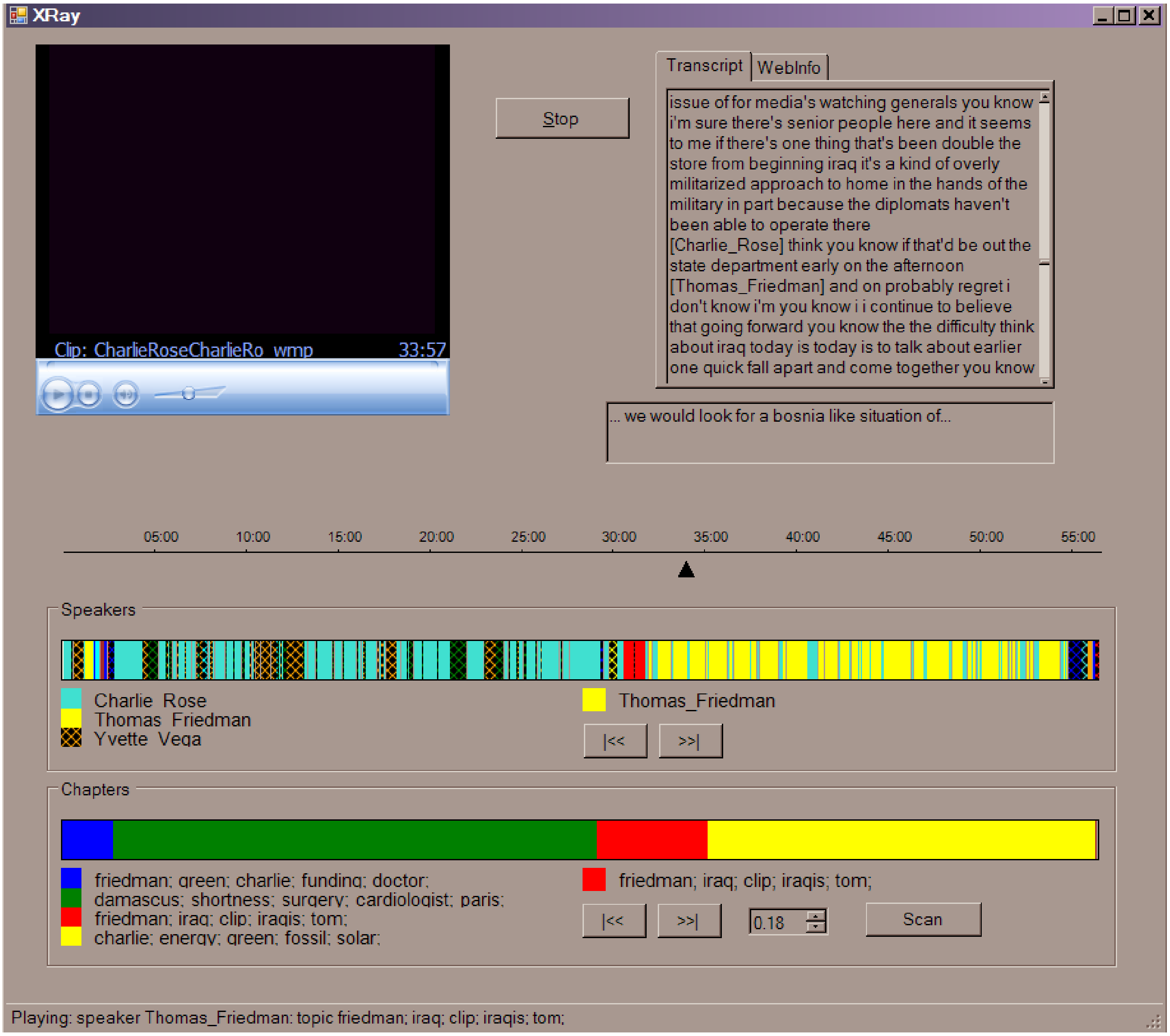

Fig. 1: Illustration for speaker and topic segmentation

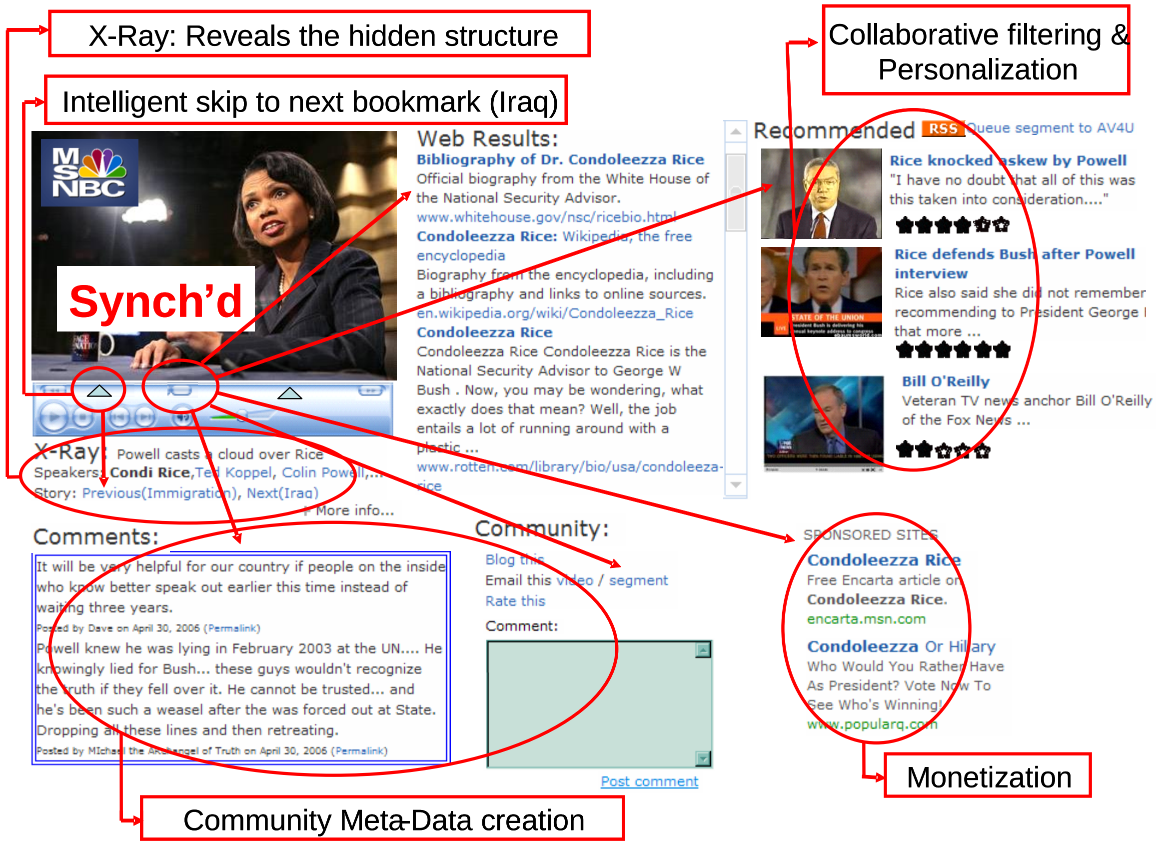

Fig. 2: Illustration of meta-data and web integration