ENHANCED VIDEO BROWSING USING AUTOMATICALLY EXTRACTED AUDIO EXCERPTS

Jonathan Foote, Matthew Cooper, and Lynn Wilcox

FX Palo Alto Laboratory
3400 Hillview Avenue Bldg. 4
Palo Alto, CA 94304 USA
{foote, cooper, wilcox}@fxpal.com

ABSTRACT

We present a method for rapidly and robustly extracting audio excerpts without the overhead of speech recognition or speaker segmentation. An immediate application is to automatically augment keyframe-based video summaries with informative audio excerpts associated with the video segments represented by the keyframes. Short audio clips combined with keyframes comprise an extremely lightweight and web-browsable interface for auditioning video or similar media, without using bandwidth-intensive streaming video or audio.

1. INTRODUCTION

The ubiquity of multimedia documents has fuelled demand for tools that support efficient multimedia browsing. Many systems support media browsing using keyframe-based techniques and interfaces. In a recent report comparing presentation summaries [1], users preferred audio-visual summaries to both text transcripts and the presentation slides. In addition, participants retained more by reviewing audio-visual summaries than from either of the other two summaries, as judged by subsequent quiz results.

In this paper, we provide a browsable audio summary comprised of audio excerpts from a source video. Although we focus here on the web-based Manga interface [2], these summarization methods can be readily applied both to other interfaces for media browsing and to general audio abstracting. As shown in Figure 1, the Manga interface provides a keyframe-based summary of digital video[1]. To construct the Manga summary, keyframes are extracted for representative video segments. Each keyframe is resized depending on the corresponding video segment's length and its novelty compared with video as a whole.

We augment the Manga's keyframes with audio clips that are selected to be both representative and to avoid truncating speech utterances. Clicking on a keyframe plays the audio excerpt, facilitating nonlinear, one-click audio browsing. Typical audio segments are short enough that they download much faster than the time required to buffer streaming video, making random-access browsing much quicker. If desired, the actual video can be replayed by double-clicking on a Manga keyframe, which initiates buffering and playback of streamed video at the corresponding point in time.

To segment the source audio, we employ the similarity analysis techniques of [3]. Similarity analysis uses the available audio to model itself; thus it requires no training and is applicable to any combination of speech, music, sound effects, and noise. The analysis produces a time-indexed novelty score in which peaks indicate audio segment boundaries. Given candidate boundaries, fitness criteria are used to select an appropriate audio excerpt for each keyframe. Additionally, the extracted audio excerpts can be combined to produce an audio summary of the video.

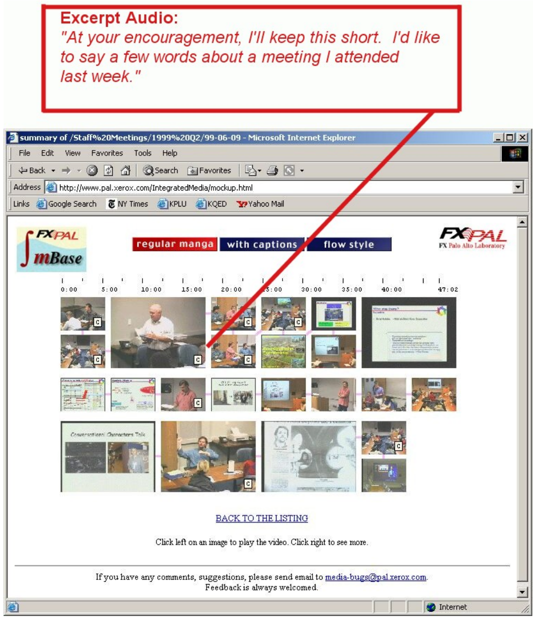
2. RELATED WORK

Previous approaches to audio segmentation and browsing can be roughly divided into computationally "light" and "heavy" approaches. Lightweight approaches primarily depend on pause detection, and thus are not robust to noise or non-speech audio. Stifelman *et al.* [4] attempt to determine spoken discourse structure using pitch and energy contours as well as a model of pause duration. "Segment beginnings" are detected automatically and used to "skim" through audio that has been associated with ink from a digital notepad. Segment lengths are not considered, nor are interfaces beyond a tape recorder metaphor.

In SpeechSkimmer [5], time compression is used to provide interactive audio summarization. Again, the speech signal is segmented according to detected pauses. The first summary mode is to remove "significant" pauses, replacing them by shorter fixed length silences to ease understanding. The other summary mode compresses the audio by deleting speech segments that are not preceded by significant (900

[1] See also http://www.fxpal.com/smartspaces/manga/

0-7803-7304-9/02/$17.00 ©2002 IEEE



Excerpt Audio:
"At your encouragement, I'll keep this short. I'd like to say a few words about a meeting I attended last week."

summary of /Staff%20Meetings/1999%20Q2/99-06-09 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.pal.xerox.com/IntegratedMedia/mockup.html
Links Google Search NY Times KPLU KQED Yahoo Mail
FXPAL mBase
regular manga
with captions
flow style
FXPAL FX Palo Alto Laboratory
0:00 5:00 10:00 15:00 20:00 25:00 30:00 35:00 40:00 47:02
Internet

BACK TO THE LISTING

Click left on an image to play the video. Click right to see more.

If you have any comments, suggestions, please send email to media-bugs@pal.xerox.com.
Feedback is always welcomed.

Fig. 1. A Manga summary of a meeting with multiple speakers and presentations. Users can click on each keyframe in the web-based interface to hear the selected audio abstract representing the corresponding segment of the meeting video. The transcript of one such excerpt is superimposed in red. Double-clicking on the keyframes opens a media player for viewing the corresponding video segment.

ms) pauses. The remaining segments are uniformly truncated. Navigation is restricted to jumping forward and backwards one segment at a time, while the Manga summary allows for random access into the multimedia file. Any nonspeech audio will seriously impact pause detection, and thus these lightweight systems are not suitable for more general audio sources such as a film or video soundtrack.

In [6], Chiu and Wilcox devised algorithms to group audio (and ink) data into useful semantic classes. Again, the audio segmentation was based on detected pauses. Segments were hierarchically clustered based on temporal distance. This method provides a valuable hierarchical browsing technique by which audio data may be organized. Slaney *et al.* have recently presented a scale-space hierarchical segmentation technique to enhance media browsing [7]. There is no summarization or time compression in either approach, but browsing is supported at varying resolutions.

Wilcox, *et al.* have used speech recognition and speech modelling to create browsable interfaces for audio [8, 9]. In this work, hidden Markov model speech segmentation methods are used to identify various acoustic classes including different speakers and music. Given the detected segments and their associated class labels, the interface allows users to navigate audio both segment by segment, and by audio class. The recorded activity is visualized by color indices indicating the audio class superimposed on timelines. Again, there is no means for summarization. Also, HMM-based techniques require training data for each audio class. The segmentation technique employed in the present system neither requires training nor makes assumptions regarding the content of the audio stream.

3. AUTOMATIC AUDIO EXCERPTING

The Manga provides a coarse level of segmentation based on color histogram analysis of the video [2]. This initial segmentation is the starting point in the following description; a single audio excerpt is selected for each of the video segments selected for the Manga interface. The combination of short audio clips with keyframes facilitates extremely lightweight and Web-browsable interface for auditioning video, without the use of bandwidth-intensive streaming video. The resulting interface might be particularly valuable for a “thin client” such as a wireless PDA.

3.1. Extracting Audio Segment Boundaries

Each video segment corresponding to a keyframe in the Manga summary is processed separately. The first step is to extract candidate audio segment boundaries *within the video segment*. For speech, we will ideally select segments containing entire phrases or sentences, but this is not generally possible without some means of speech understanding. In very limited domains such as news broadcasts where the speakers are known in advance, then speech can be segmented via statistical speech recognition or speaker segmentation [8].

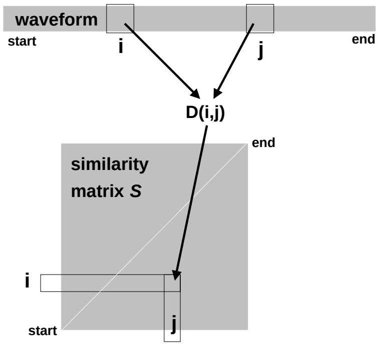


Fig. 2. Diagram of the similarity matrix embedding.

To avoid restrictive assumptions regarding the source audio, we apply self-similarity analysis for audio segmentation following [3]. We first window the digital audio at 10 Hz and compute mel-frequency cepstral coefficients. The coefficient feature vectors are compared via the cosine similarity measure. We embed this pairwise similarity data in the similarity matrix, as illustrated in Figure 2. We then calculate the novelty score via kernel correlation. Peaks in the novelty score are automatically detected and labelled as audio segment boundaries. Maxima in the novelty score are excellent candidate segment boundaries, because the audio will be self-similar between maxima and significantly different across them. Additionally, unlike the aforementioned speech recognition and speaker segmentation approaches, similarity analysis requires no training, substantially less computation, and is applicable to general audio sources including music. This technique also has many advantages over pause detection; foremost it will work even in significantly noisy conditions or with background music present, conditions that will cause silence-detection methods to fail. Because of the non-linear distance measure, the novelty score will still produce sharp and distinct peaks even when calculated over a large time extent. Methods using average energy will result in less distinct peaks for longer averaging windows, resulting in decreased time resolution.

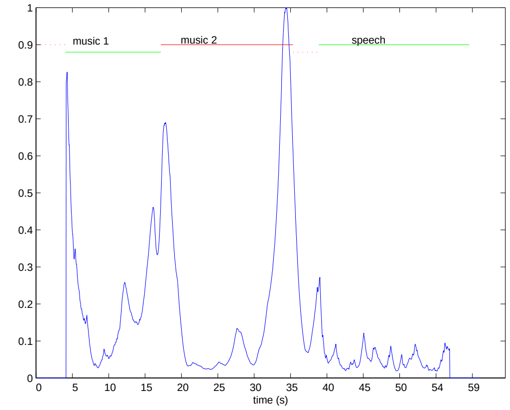


Fig. 3. Novelty score for audio comprised of speech and music.

Figure 3 shows the audio novelty for the first minute of Animals Have Young (video V14 from [10]). This segment contains 4 seconds of introductory silence, followed by a short musical segment with the production logo. At 17 seconds the titles start, and very different theme music commences. At 35 seconds, this fades into a short silence, followed by female speech over attenuated background music for the remainder of the segment. The novelty score is computed over a 8-second window to average out the short-time spectral differences in speech. The largest peak occurs directly on the speech/music transition at 35 seconds. The two other major peaks occur at the transitions between silence and music at 4 seconds and between the introduction and theme music at 17 seconds. Using a naive spectral distance measure, the novelty score can't in general, discriminate between speakers unless they have markedly different vocal spectra (such as between genders). As self-similarity analysis has been shown to be a promising technique for segmenting both audio [3] and video [11], we also envision audio segment selection based on joint analysis of the audio and video novelty scores.

3.2. Determining Candidate Segments

The novelty score produces a list of candidate audio start and end points. The objective is to select meaningful segments without truncating speech utterances and to avoid segments consisting mostly of silence. We have experimented with several heuristics to select the representative audio excerpt using segment length and average segment energy.

Figure 4 shows a flowchart of the algorithm. Of all the detected peaks, the N highest are chosen. For the current embodiment, N is half the number of seconds in the audio to be analyzed, resulting in an average of one peak every two seconds. Every selected peak is considered to be the start boundary of a candidate segment. We wish to satisfy a length constraint, such that selected segments are neither too short nor too long. Thus for every start boundary, we consider all successive peaks within a range of between 5 and 15 seconds as candidate end boundaries. There may be more than one peak within this range, thus we consider multiple segments with the same start point but different end points. Every possible start/end combination satisfying the length constraint is thus considered; because of the short lengths, in practice, this is not excessive. Each possible segment is added to a list, and a fitness score is computed for each, as detailed in the following section.

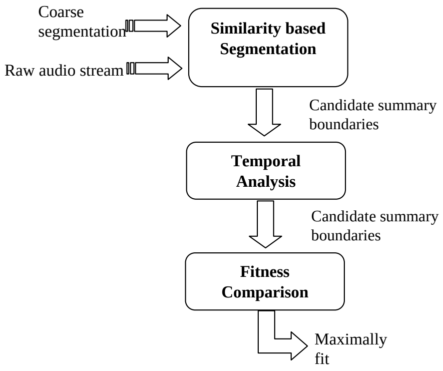

Coarse segmentation
Raw audio stream
Similarity based Segmentation
Candidate summary boundaries
Temporal Analysis
Candidate summary boundaries
Fitness Comparison
Maximally fit

Fig. 4. Flowchart describing summarization method.

3.3. Segment fitness measures

To determine the fitness score, we first compute the average energy of each candidate audio segment. We also calculate an additional penalty to favor segments which follow lengthy silences and are thus more likely to be complete or initial utterances: the average signal energy of the second immediately preceding the audio segment. The fitness score is simply the average segment energy minus the average energy of the second of audio *preceding* the segment. The audio excerpt that satisfies the length constraint and maximizes the fitness score is then selected.

Many more sophisticated fitness measures can be used. Segments can be ranked by acoustic similarity to speech or other audio, [12, 13]. This avoids the problem of energy-based methods, where high-energy events such as laughter or applause are chosen in preference to speech utterances. If the segments are to be linked to video keyframes, another enhancement is to constrain the audio excerpt interval to contain the keyframe occurrence time; that is the keyframe should occur sometime during the selected segment. Any other source of time-stamped metadata can also be used to select segments, such as annotations, captions or subtitles (perhaps based on keyword or tf/idf measures), or any other information source.

4. EXPERIMENTS

At FXPAL, we maintain a multimedia database of recorded staff meetings and seminars which can be accessed via the Manga system. We processed several recorded staff meetings to attach audio excerpts to the keyframes, and we built a prototype extension to the Manga interface, shown in Figure 1. The keyframes permit users to quickly locate particularly interesting sections of the meeting videos. Previously, single-clicking a keyframe would play the video segment, often with the audio starting in the middle of a word or sentence and creating confusion. We now attach an audio excerpt to each video segment, using the excerpting techniques described above. The excerpt is played when the keyframe is single-clicked, allowing more detailed browsing without the overhead and bandwidth requirements of actual video. The video segment can be viewed by double-clicking on the keyframe.

5. CONCLUSION

In this paper we have presented a lightweight framework for browsing video using keyframes and audio excerpts. The interface is based upon the Manga system, and uses its video segmentation as a starting point for selecting representative audio segments. An audio excerpt is attached to each video keyframe providing an enhanced means by which users can quickly browse video. The techniques employed are applicable to general audio and video content.

6. ACKNOWLEDGMENT

The authors acknowledge Hari Sundaram's contributions in preliminary experiments with these ideas.

7. REFERENCES

[1] L. He, E. Sanocki, A. Gupta, J. Grudin. Comparing Presentation Summaries: Slides vs. Reading vs. Listening. *Microsoft Research Tech. Rep. MSR-TR-99-68*, 1999.

[2] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video Manga: Generating Semantically Meaningful Video Summaries. *Proc. ACM Multimedia*, pp. 383-392, 1999.

[3] J. Foote, Automatic Audio Segmentation using a Measure of Audio Novelty. *Proc. IEEE ICME* **1**:452-455, 2000.

[4] L.J. Stifelman. *The Audio Notebook: Paper and Pen Interaction with Structured Speech*. Ph.D. Thesis, MIT, 1997.

[5] B. Arons. Pitch Based Emphasis Detection for Segmenting Speech Recordings, *Proc. ICSLP 1994* **4**:1931-1934, 1994.

[6] P. Chiu and L. Wilcox. A Dynamic Grouping Technique for Ink and Audio Notes. *Proc. ACM UIST*, pp. 195-202, 1998.

[7] M. Slaney, D. Ponceleon, and J. Kaufman. Multimedia Edges: Finding Hierarchy in all Dimensions. *Proc. ACM Multimedia*, 2001.

[8] L. Wilcox, F. Chen, and V. Balasubramanian. Segmentation of speech using speaker identication. *Proc. IEEE ICASSP 1994*, **S1**:161–164, 1994.

[9] D. Kimber and L. Wilcox. Acoustic Segmentation for Audio Browsers. *Proc. Interface Conference*, July 1996.

[10] MPEG Requirements Group, Description of MPEG-7 Content Set, 1998.

[11] M. Cooper and J. Foote. Video Segmentation via Self-Similarity Analysis. *Proc. IEEE ICIP*:378-81, 2001.

[12] J. Foote. Content-based Retrieval of Music and Audio. *Proc. SPIE* **3229**:138-147, 1997.

[13] S. E. Johnson and P. C. Woodland. A Method for Direct Audio Search with Applicactions to Indexing and Retrieval. *Proc. IEEE ICASSP*, 2000.