Comparing Presentation Summaries:
Slides vs. Reading vs. Listening

Liwei He, Elizabeth Sanocki, Anoop Gupta, Jonathan Grudin

September 20, 1999

Technical Report
MSR-TR-99-68

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052

Comparing Presentation Summaries: Slides vs. Reading vs. Listening

Liwei He, Elizabeth Sanocki, Anoop Gupta, Jonathan Grudin
Microsoft Research
One Microsoft Way, Redmond, WA 98052
+1 (425) 703-6259
{lhe,a-elisan,anoop,jgrudin}@microsoft.com

ABSTRACT

As more audio and video technical presentations go online, it becomes imperative to give users effective summarizing and skimming tools so that they can find the presentation they want and browse through it quickly. In a previous study we reported various automated methods for summarizing audio-video of presentations, and user response. An open question remained about how well various text/image only techniques will compare to the audio-video summarizations. This study attempts to fill that gap.

This paper reports a user study that compares four possible ways of allowing a user to skim a presentation: 1) PowerPoint slides used by the speaker during the presentation, 2) the text transcript created by professional transcribers from the presentation, 3) the transcript with important points highlighted by the speaker, and 4) a audio-video summary created by the speaker. Results show that although some text-only conditions can match the audio-video summary, users have a preference for audio-video. Furthermore, different styles of slide-authoring (e.g., detailed vs. big-points only) can have a big impact on their effectiveness as summaries, raising a dilemma for some speakers in authoring for on-demand previewing versus that for live audiences.

Keywords

Video abstraction, summarization, evaluation, digital video library, video browsing, video skim, multimedia, transcript

INTRODUCTION

Digital multimedia content is becoming pervasive both on corporate intranets and on the Internet. Many corporations are making audio and video of internal seminars available online for both live and on-demand viewing, and many academic institutions are making lecture videos and seminars available online. For example, research seminars from Stanford, Xerox PARC, University of Washington, and other sites can be watched at the MURL Seminar Site (http://murl.microsoft.com). Microsoft's corporate intranet has hundreds of presentations available on it, and close to 10,000 employees have watched one or more presentation [7]. These numbers are likely to grow dramatically in the near future. With thousands of hours of such content available on-demand, it becomes imperative to give users necessary summarizing and skimming tools so that they can find the content they want and browse through it quickly.

One solution technique that can help in browsing is *time compression* [3,11]. It allows the *complete* audio-video to be watched in a shorter amount of time by speeding up the playback with no pitch distortion. This technique, however, allows only a maximum time saving of a factor of 1.5-2.5 depending on the speech speed [3], beyond which the speech starts to become incomprehensible. Further increase in compression ratio is possible [5], but at cost of increasing the software complexity and listeners' concentration and stress level. People tend to feel most comfortable with a rate of about 1.4 [e.g.,11].

Getting a much higher factor of time-savings (factors of 3-10) requires creating an audio-video summary of the presentation. A summary by definition implies that portions of the content are thrown away. For example, we may select only the first 30 seconds of audio-video after each of the slide transitions in the presentation, or have a human identify key portions of the talk and include only those segments, or base it on the access patterns of users who have watched the talk before us.

In an earlier paper [8], we studied three automatic methods for creating audio-video summaries for presentations with slides. These were compared to author-generated summaries. While users preferred author-generated summaries, as may be expected, they showed good comprehension with automated summaries and were overall quite positive about automated methods. The study reported in this paper extends our earlier work by experimenting with non-video summarization abstractions to address the following questions:

- Since all of the audio-video summaries included slides, how much of the performance/comprehension increment was due to slides alone? In fact, this is the most common way in which presentation are archived on the web today---people simply post their slides. What is gained by skimming just the slides?
- How will people perform with the *full text transcripts* of the presentation, in contrast to the audio-video summaries? Two factors motivate this. First, speech-

to-text technology is getting good enough that this may become feasible in the not-so-distant future. Second, people are great at skimming text to discover relevance and key points. Perhaps given a fixed time to browse the presentation, they can gain more from skimming a full text transcript than spending the same time on an audio-video summary.

- If we highlight the parts of the transcript that a speaker included in a video summary, would performance be comparable to or better than the performance with the video summary? The highlighted transcript and the video summary would each provide the information that a speaker thinks is important. Would users prefer skimming the text transcript or watching the audio-video summary?

These questions motivated the study presented below. We compare four conditions: slides-only, full text-transcript with no highlights, full text-transcript with highlights, and audio-video summary. We also compare to the results to the earlier study. We find that although full text transcript with highlights condition can match the audio-video summary, users have a slight preference for audio-video. Furthermore, different styles of slide-authoring (e.g., detailed vs. big-points only) can have a big impact on their effectiveness as summaries, raising a dilemma for some speakers in authoring for on-demand previewing versus that for live audiences.

The paper is organized as follows: The next section describes the previous work on automatic summarization that this study extends. Next, the experimental design of the current study is presented, followed by the results section. Finally, we discuss related work and draw conclusions.

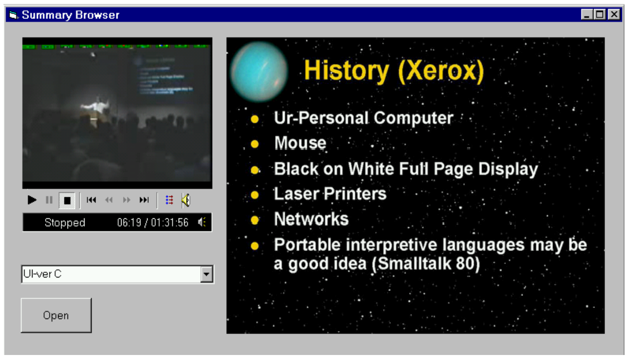
AUTOMATIC AUDIO-VIDEO SUMMARIZATION

We briefly summarize our earlier study on automated audio-video summarization methods [8]. The combination of the current study and this older study enable us to build a more complete picture of the overall tradeoffs.

Our study used a combination of information sources in talks to determine segments to be included in the summary. These were: 1) analysis of speech signal, for example, analysis of pitch, pauses, loudness over time; 2) slide-transition points, i.e., where the speaker switched slides; and 3) information about other users access patterns (we used details logs indicating segments that were watched or skipped by previous viewers).

We experimented with three algorithms based on these sources of information: i) *slide-transition* points only (S); ii) identification of emphasized speech by *pitch activity* analysis (P), using an algorithm introduced by Arons [3]; iii) a combination of slide transitions, pitch activity, and previous user access patterns (SPU). In addition, we obtained a human-generated video summary (A) by asking the author-instructor for the talk to highlight segments of transcript[1].

For our study, four presentations were obtained from an internal training web site. Each author was given the text transcript of the talk with slide transition points marked. They marked summary segments with a highlighting pen. These sections were then assembled into a video summary by aligning the highlighted sentences with the corresponding video. A study of 24 subjects was then conducted to compare the summaries created by the authors to the three automatically generated summaries.



History (Xerox)

- Ur-Personal Computer
- Mouse
- Black on White Full Page Display
- Laser Printers
- Networks
- Portable interpretive languages may be a good idea (Smalltalk 80)

Figure 1: The interface for the experimental software.

Figure 1 shows the display seen by subjects watching the summaries. All video summaries are shown with the associated slides. As the video window, shown in the upper-left, plays the segments in the summary, the slides in the right pane change in synchrony.

We used two measures: performance improvement on quizzes before and after watching the video summary, and ratings on an opinion survey.

The outcome for the first measure was that author-generated summaries resulted in significantly greater improvement than computer-generated summaries (95% confidence level). The automated methods also resulted in substantial improvement (results presented later in this paper), but they were statistically indistinguishable from each other.

One hypothesis for lack of significant difference between the automated methods was that most of the useful information may come from the slides. Although the audio-video segments selected for summary were quite different for the different methods, the slides shown were substantially the same (as slide transitions are very infrequent). However, participants estimated that slides carried only 46% of the information and audio-video carried 54%. So the hypothesis is not quite justified; current study could help.

Survey responses also indicated a preference for author-generated summaries. The preference was greater along

[1] In one case, the author was unavailable and designated another expert to highlight the summary.

some dimensions (e.g., author-generated summaries were judged to be much more coherent), while along other dimensions (e.g., confidence that key points were covered by the summary) the author-generated and automated summaries did comparably.

Overall, the computer-generated summaries were well received by participants, many of whom expressed surprise upon being told afterwards that a computer generated them.

NEW SUMMARIZATION ABSTRACTIONS

In this study we extend the previous work by examining three non-video summarizations or abstractions: slides only (SO), text transcripts with slides (T), transcripts highlighted by the authors with slides (TH). Author-generated video summaries with slides (A), which are the same used in the earlier study, are included to provide a comparison with the results of previous study.

Slides Only (SO)

Technical presentations are usually accompanied with slides that set the context for the audience, indicating what was just said and what will be addressed next. Speakers also use the slides as cues for themselves. Normally, much of presentation preparation goes into preparing the slides: deciding how many slides, which ideas go onto which slides, and so forth. Because so much energy is put into the slides, it seems natural to use them in a summary whenever possible. Furthermore, slides is what people frequently post on the web, slides is what they send around in email, so it is useful to understand how well people comprehend just using slides.

Text Transcript with Slides (T)

People are great at skimming text to discover relevance and key points. Perhaps given a fixed time to browse the presentation, they can gain more from skimming a full text transcript than spending the same time on an audio-video summary. Text transcripts are also interesting because commercial dictation software, such as ViaVoice from IBM and NaturallySpeaking from Dragon Systems, can produce text transcript automatically. The error rates are high without training, but close to 5% with proper training and recording condition. Speech-to-text will continue to improve and may become feasible for lecture transcription in the not-so-distant future.

For this condition we assumed the ideal case, and had all of the presentations fully transcribed by human. We then manually segment the text into one paragraph per slide. The title of the slide is also inserted in front of each paragraph. The process can be made fully automatic if we later use speech-to-text software, which gives the timing information of the text output, and have the slide transition times. The slides were also made available to the subjects in this condition.

Transcript with Key Points Highlighted and Slides (TH)

The benefit of providing the full text transcript is that every word that was said during the presentation is captured. The disadvantage is that the transcript is a written form of spoken language, which contains filler words, phrases, and repetitions. It can be longer and harder to read than a paper or a book that is written specifically for reading and has the formatting and structuring elements to assist reading and skimming.

Viewers could benefit from having key parts highlighted. Our first study showed that automatic summarization techniques have much room for improvement, so again we chose to use the ideal case, the transcript highlighted by an author or expert. Each author was given the text transcript of the talk with slide transition points marked. They marked summary segments with a highlighting pen. The same sections were also assembled into the video summary (A) by aligning the highlighted sentences with the corresponding video. The highlighted parts are presented to the subjects as bold and underlined text on screen. Again, slides were also made available to the subjects in this condition.

Talks Used in the Study

We reused the four presentations and quizzes from the previous study to permit comparison of results. The talks were on the topics of user-interface design (UI), Dynamic HTML (DH), Internet Explorer 5.0 (IE), and Microsoft Transaction Server (MT).

Table 1: Information associated with each presentation.

	UI	DH	IE	MT
Duration (mm:ss)	71:59	40:32	47:01	71:03
# of slides	17	18	27	52
# of slides / min	0.2	0.4	0.6	0.7
# of words	15229	8081	6760	11578
% of words highlighted	19	24	25	20
Duaration of AV summary (mm:ss)	13:44	9:59	11:37	14:20

Table 1 shows some general information associated with each talk. It is interesting to note the wide disparity in number of slides associated with each talk. For example, although UI and MT are both around 70 minutes long, one has 17 slides and the other 52. Also note that the fraction of words highlighted by the speaker in the summaries is about 20-25%. Obviously, the end results may be different if much less or much higher summarization factors were chosen. A factor of 4-5 summarization seemed an interesting middle ground to us.

EXPERIMENTAL DESIGN

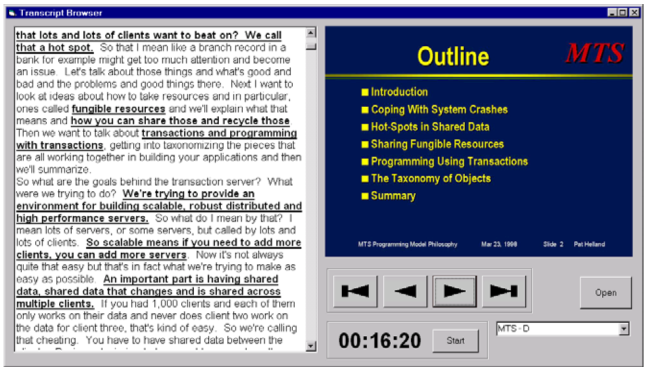
The same measures were taken as in the first study: quizzes on objective learning and surveys to gauge subjective reactions.

Each presentation author had written 9 to 15 quiz questions that required participants to draw inferences from the content of the summary or to relay factual information contained in the summary. We selected 8 from each to construct a 32-question multiple-choice test.

The 24 participants were employees and contingent staff members of a software company working in technical job positions. All lacked expertise in these four topic areas. Participants were given a gratuity upon completing the tasks.

Participants first completed a background survey and took the quiz to document their initial knowledge level. We randomly ordered questions within and across talks, so that people would have less ability to be guided by questions while watching the talk summaries.

Each participant watched or read four summaries, one for each talk and one with each summarization technique. Talk order and summarization technique were counterbalanced to control for order effects.



that lots and lots of clients want to beat on? We call that a hot spot. So that I mean like a branch record in a bank for example might get too much attention and become an issue. Let's talk about those things and what's good and bad and the problems and good things there. Next I want to look at ideas about how to take resources and in particular, ones called **fungible resources** and we'll explain what that means and **how you can share those and recycle those**. Then we want to talk about **transactions and programming with transactions**, getting into taxonomizing the pieces that are all working together in building your applications and then we'll summarize.

So what are the goals behind the transaction server? What were we trying to do? **We're trying to provide an environment for building scalable, robust distributed and high performance servers.** So what do I mean by that? I mean lots of servers, or some servers, but called by lots and lots of clients. **So scalable means if you need to add more clients, you can add more servers**. Now it's not always quite that easy but that's in fact what we're trying to make as easy as possible. **An important part is having shared data, shared data that changes and is shared across multiple clients.** If you had 1,000 clients and each of them only works on their data and never does client two work on the data for client three, that's kind of easy. So we're calling that cheating. You have to have shared data between the

Outline

MTS

- Introduction
- Coping With System Crashes
- Hot-Spots in Shared Data
- Sharing Fungible Resources
- Programming Using Transactions
- The Taxonomy of Objects
- Summary

MTS Programming Model Philosophy Mar 23, 1998 Slide 2 Pat Helland

Figure 2: Interface for the conditions SO, T, and TH. The participant can use the vertical scroll bar to navigate the text transcript or use the four control buttons (shown below the slides) to navigate the slides. However, the current slide and the displayed text transcript are not linked. This allows the participant to view the slide in one part and review the transcript in another area. The countdown timer below the slide-navigation controls serves as a reminder of how much time left to review the current summary.

The display for video summary condition (A) was the same as for our previous study (see Figure 1). Figure 2 shows the interface for the other three conditions. In the slide-only condition, the left transcript pane is blank. While watching or reading a summary, a participant was given the same time as the duration of the audio-video summary of corresponding talk (see Table 1). They were free to navigate within the slides and transcript. Once finished, however, participants were instructed not to review portions of the summary. Participants were provided pen and paper to take notes. After each summary, participants filled out the subjective survey and retook the quiz.

RESULTS

Evaluating summarization algorithms is a fundamentally difficult task, as the critical attributes are highly complex and difficult to quantify computationally. We use a combination of performance on a quiz and ratings on an opinion survey for our evaluation.

Quiz Results

We expected the author-generated summaries (TH and A) to produce the highest quiz scores, as the quizzes were created by the authors. However, we wanted to know: i) Are there significant differences between the author-generated video summary and the text transcript with the same portion highlighted? ii) How much worse are SO and T compared to A and TH? iii) Are there performance differences across the talks?

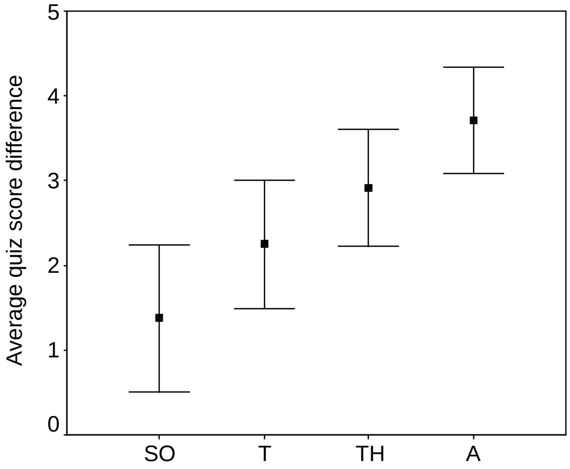


Figure 3: Quiz score improvement by condition with 95% confidence intervals. The apparent linearity of the quiz score improvement is coincidental.

Figure 3 shows the difference between pre-summary and post-summary quiz scores as a function of the conditions. Quiz scores were improved most by the audio-video summaries (A). To a lesser extent, quiz scores were improved by the summaries that combined highlighted transcripts and slides (TH). The smallest improvements were obtained from the slides alone (SO) and transcript with slides (T) versions.

The above data show quiz scores with audio-video summary (A) are significantly better than SO and T. When presenting our previous study's results to audiences, some have suggested that just providing the text-transcript should be adequate. This study shows that there are significant differences. The quiz scores for A and TH did not differ significantly (at level .05), but they are significantly different at level 0.07. It appears there is significant value added from hearing the speaker's voice and intonation. We present some intuition regarding user's preference for audio summaries later in this section.

In terms of the cost of production, SO costs the least and A costs the most, so the amount of score improvements does correlate with the amount of effort needed to produce the summaries. On a first glance, this suggests that the more effort that goes into producing the summary the better the improvement. But let us examine closer and separate the quiz scores for individual talks.

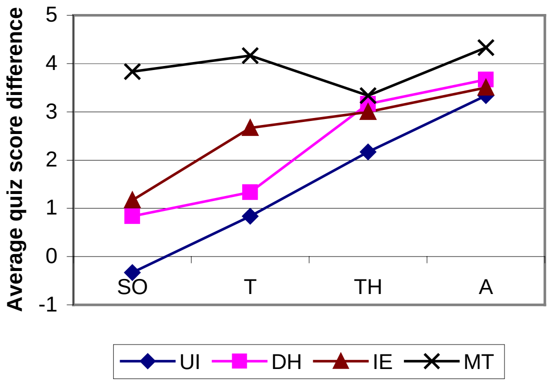


Figure 4: Average difference in quiz score by summary versions.

Figure 4 shows the average quiz score improvement by summary versions for each talk. The quiz score improvement for the video summary condition (A) is highest and least varied across talks. In contrast, the slides-only quiz score improvements (SO) were highly variable.

The amount of variability in the quiz score improvements seemed to correlate with the extent of information present in the slides. For example, one measure is the number of slides per minute.[2] By this metric (see Table 1) the talks are ordered as MT highest (0.7 slides/min), IE (0.6 slides/min), DH (0.4 slides/min) and UI (0.2 slides/min). As predicted, we see that the variance as a function of summarization method is least for MT talk and most for UI talk. Our intuition regarding this is that when the information is not present in the slides (the base condition) then the summarization method affects comprehension much more.

Survey Results

Participants completed a short survey after watching each summary. The surveys were administered prior to repeating the quiz so that quiz performance would not affect their opinions on the surveys.

User Ratings

The pattern of responses was similar to that of the quiz scores (see Table 2 and Table 3). Ratings for the video summaries (A) tended to be the highest. However, they were not significantly different from the ratings for the highlighted transcript (TH). Some of the ratings (synopsis, efficient) for A and TH as a group were significantly greater than those for the slides-only (SO) and transcript (T). Some other ratings (confidence, skip) SO were significantly lower than the other three.

Table 2: Post-quiz survey results by conditions[3].

By condition	Synop.	Effi.	Enjoy	Key points (%)	Skip talk	Concise	Cohe.
A	4.96	5.04	4.78	68.91	4.41	5.13	4.13
SO	3.13	3.38	3.33	41.25	1.96	2.92	2.83
T	3.58	3.25	3.29	61.67	3.83	3.50	4.17
TH	4.70	4.61	3.83	64.13	4.52	4.52	4.35

Also following the quiz score trend is the fact that the ratings for the MT talk were higher than the others (see Table 3). The MT talk was well liked among the participants. It was rated consistently higher, independent of summary method. Again this is probably due to the fact that the slides were sufficiently detailed so that they could "stand alone" and be interpreted without the speaker present.

Table 3: Post-quiz survey results by talks.

By talk	Synop.	Effi.	Enjoy	Key points (%)	Skip talk	Concise	Cohe.
UI	3.79	3.96	3.92	52.50	3.04	3.71	3.88
DH	4.17	3.96	3.74	60.22	3.78	3.91	3.73
IE	3.83	3.96	3.30	51.09	3.36	3.83	3.96
MT	4.50	4.33	4.21	71.25	4.42	4.54	4.61

User Comments

MT talk aside, most of the participants found that the slides only condition (SO) lacked sufficient information. They also felt scanning the full text in condition T tedious. Thirteen of the 24 participants rated the audio-video summary (A) as their favorite summary abstraction, while eleven chose the highlighted transcript with slides (TH).

Participants liking the audio-video summary did so mainly because it was more passive, self-contained, and multi-modal. One participant said, "It felt like you were at the presentation. You could hear the speaker's emphasis and inflections upon what was important. It was much easier to listen and read slides versus reading transcripts and

[2] Of course, this does not take into account the amount of information within each slide.

[3] Complete wording: 1) Synopsis: "I feel that the condition gave an excellent synopsis of the talk." 2) Efficient: "I feel that the condition is an efficient way to summarize talks." 3) Enjoyed: "I enjoyed reading through (or watching) the condition to get my information." 4) Key points: "My confidence that I was presented with the key points of the condition is:" 5) Skip talk: "I feel that I could skip the full-length video-taped talk because I read (or watch) the condition." 6) Concise: "I feel that the condition captured the essence of the video-taped talk in a concise manner." 7) Coherent: "I feel that the condition was coherent–it provided reasonable context, transitions, and sentence flow so that the points of the talk were understandable." Responses were from 1 ("strongly disagree") to 7 ("strongly agree").

reading slides." Another commented, "It kept my interest high. It is more enjoyable listening and seeing the presenter."

Participants liking the highlighted transcript with slides condition most did so because it gave them more control over the pace and allowed them to read what they considered important. One participant liking the highlighted transcript most commented, "I felt this was a more efficient way to get a summary of the presentation. … I could re-read the portions I was interested in or unclear about." Another said, "I like having the option of being able to get more detailed info when I need it."

Comparison with the Automatic Summary Study

There are several similarities between this study and our previous study on automatic summary algorithms: i) The talks and quiz questions were the same; ii) The author-generated audio-video summary (condition A) was present in both studies; iii) Slides were shown in all conditions in both studies; and iv) The studies evaluated performance using quizzes and ratings on opinion surveys. Given these similarities, we can compare the results from these two studies.

Figure 5 shows the average quiz score difference by conditions from the automatic summary study. Compared with Figure 4, there is no clear correlation between the variability among the talks and conditions. It may be because the differences between the computer-generated video summaries are not as big as the differences between conditions S, T, and TH.

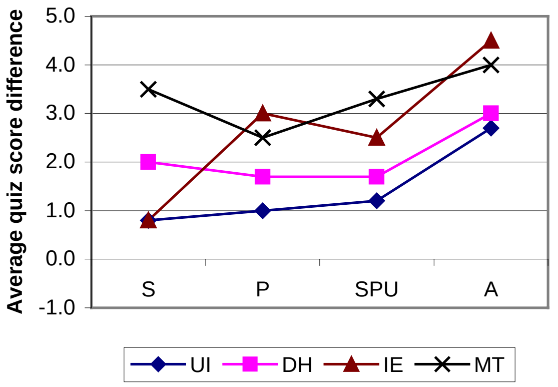


Figure 5: Average quiz score difference by conditions from the automatic summary study.

In Table 4, we list the post-quiz ratings that are in common between the two studies. The top half of the table shows the ratings for the previous study, while the bottom half shows the ratings for this study.

Condition A was included in both studies, though its ratings were consistently lower in the present study. One hypothesis is that the ratings are relative to the quality of other conditions in the same study. The author-generated audio-video summaries were higher in quality than were the summaries generated by the computer. In the current study, the audio-video summaries were comparable in quality to the transcripts with highlighted summary (TH).

The slide-transition-based summary in the previous study (Condition S) assembled a summary by allocating time to each slide in proportion to the amount of time that the speaker spent on it in the full-length talk. Thus condition S differed from the slide-only condition (SO) in the present study by showing audio-video in addition to all the slides of the talk. From Table 4 we see that the ratings for condition S are consistently higher than condition SO, suggesting that providing an audio-video summary can add a lot of value to the slides, even when the summary is created with a simple summarization technique.

Table 4: Responses to quality of summary for various methods for the automatic summary study (top half) and the current study (bottom half).

	Synopsis	Key points (%)	Skip talk	Concise	Coherent
SPU	4.92	64.17	3.54	4.63	3.58
P	4.83	62.50	3.04	4.13	3.46
S	4.33	56.25	3.21	4.08	3.57
A*	5.00	76.25	4.96	5.63	5.33
A	4.96	68.91	4.41	5.13	4.13
SO	3.13	41.25	1.96	2.92	2.83
T	3.58	61.67	3.83	3.50	4.17
TH	4.70	64.13	4.52	4.52	4.35

* The data for A here is from our previous study on automatic summarization algorithms.

One surprising result in the previous study was that participants rated the computer-generated summaries more positively as they progressed through the study. The summary shown to the participants last in each session was consistently rated as being clearer (p=.048), less choppy (p=.001), and of higher quality (p=.013) than were the first three summaries in the same session independent of condition.. The study was designed so that each of the four summary methods was presented equally often in each position in the sequence. We found no such effect in the current study. However, summary presentation styles varied more in the current study, possibly reducing the chance for the participants to habituate to disadvantages of each abstraction.

DISCUSSION AND RELATED WORK

There has been considerable research on indexing, searching and browsing the rapidly expanding sources of digital video [1,2,5,9,10,12,14,17,18]. These approaches all focus on automatic techniques based on visual and aspects of media, primarily employing image-recognition and image-processing techniques. Some of them [10,14] use textual information from speech-to-text software or closed captions. Our study complements these systems by providing a user study that proved the usefulness of audio-video browsing and summary system.

Christel et al. [4] report a subjective evaluation of summaries created from image analysis, keyword speech recognition, and combinations, again from general-purpose video. Based on analysis, summaries or skims are constructed from 3-5 second video shots. They tested the quality of skims using image recognition and text-phrase recognition tasks. Performance and subjective satisfaction of all skimming approaches contrasted unfavorably with viewing the full video; satisfaction was less for each technique on each dimension examined. Our study extends this paper by comparing additional non-video summarization abstractions with video summaries.

Barry Arons' SpeechSkimmer [3] allows audio to be played at multiple levels of detail. Speech content can be played at normal speeds, with pauses removed, or restricted to phases emphasized by the speaker. A knob orthogonally controls pitch-preserved time-compression of the speech. Lisa Stifelman introduced Audio Notebook, a prototype note-pad combining pen-and-paper and audio recording [15,16]. Audio Notebook relies on the synchronization of key points marked by pen on paper to structure the recorded audio. These two systems provide ways to skim the audio – we can use these systems to replace the transcript-browsing interface in our summary system.

CONCLUDING REMARKS

As storage cost drops, network bandwidth increases, and inexpensive video cameras becomes available, more audio and video technical presentations will go online. Given this expected explosion, it becomes imperative to give users effective summarizing and skimming tools so that they can find the presentation they want and browse through it quickly.

This paper reports a study that extends our previous work by comparing three non-video summarization abstractions with an audio-video summary created by the speaker. The three non-video summary techniques are: 1) PowerPoint slides in the presentation, 2) a text transcript created from the presentation, and 3) the transcript with important points highlighted by the speaker.

Results show that although transcripts-with-highlights condition can match the audio-video summary, users have a preference for audio-video. Slides-only and plain transcripts are significantly worse than audio-video summaries. Furthermore, different styles of slide-authoring (e.g., detailed vs. big-points only) can have a big impact on their effectiveness as summaries. The result contradicts the common advice for creating succinct slides when giving talks to live audiences. This raises a dilemma for speakers who are authoring for both on-demand and live audiences. On solution might be to create two versions of slides. The succinct version can be used in the live presentation, while the more detailed version is placed online.

The two-versions of slides solution, of course, requires cooperation from the authors. As the technology for creating computer-generated summaries improves, the amount of author work in the creation of summaries should be reduced. At the same time, as more people browse audio-video online, authors may often be more willing to contribute to improving their experience. An interesting future direction is technology-assisted tools that allow authors to very quickly indicate important segments (e.g., speech-to-text transcript marked by author in 5 minutes using a tool).

ACKNOWLEDGMENT

Thanks to the Microsoft Usability Labs for use of their lab facilities. Steve Capps, Pat Helland, Dave Massy, and Briand Sanderson gave their valuable time to create the summaries and quiz questions for their presentations. Gayna Williams and JJ Cadiz reviewed the paper and gave us valuable suggestions for improvement.

REFERENCES

1. Aoki, H., Shimotsuji, S. & Hori, O. A Shot Classification Method of Selecting Effective Key-frames for Video Browsing. In Proceedings of the 6th ACM international conference on Multimedia, 1996, pp 1-10.
2. Arman, F., Depommier, R., Hsu, A. & Chiu M.Y. Content-based Browsing of Video Sequences, In Proceedings of the 6th ACM international conference on Multimedia, 1994, pp 97-103.
3. Arons, B. SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM Transactions on Computer Human Interaction, 4,* 1, 1997, 3-38.
4. Christel, M.G., Smith, M.A., Taylor, C.R. & Winkler, D.B. Evolving Video Skims into Useful Multimedia Abstractions. In *Proceedings of CHI, April 1998,* pp. 171-178.
5. Covell, M., Withgott, M., & Slaney, M. Mach1: Nonuniform Time-Scale Modification of Speech. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle WA, May 12-15 1998.
6. Foote, J., Boreczky, J., Girgensohn, A. & Wilcox, L. An Intelligent Media Brower using Automatic Multimodal Analysis. In *Proceedings of ACM Multimedia, September 1998*, pp. 375-380.
7. He, L., Gupta, A., White, S.A., & Grudin, J., 1999. Design lessons from deployment of on-demand video. *CHI'99 Extended Abstracts*, 276-277. ACM.
8. He, L., Sanocki, E., Gupta, A., & Grudin, J., 1999. Auto-summarization of audio-video presentations. In *Proc. Multimedia'99.* ACM.
9. Lienhart, R., Pfeiffer, S., Fischer S. & Effelsberg, W. Video Abstracting, *ACM Communications*, December 1997.
10. Merlino, A., Morey, D. & Maybury, M. Broadcast News Navigation Using Story Segmentation. In Proceedings of the 6th ACM international conference on Multimedia, 1997.
11. Omoigui, N., He, L., Gupta, A., Grudin, J. & Sanocki, E. Time-compression: System Concerns, Usage, and Benefits. Proceedings of *ACM Conference on Computer Human Interaction,* 1999.

12. Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D. & Diklic, D. Key to Effective Video Retrieval: Effective Cataloging and Browsing. In Proceedings of the 6th ACM international conference on Multimedia, September 1998.

13. Stanford Online: Masters in Electrical Engineering, 1998. http://scpd.stanford.edu/cee/telecom/onlinedegree.html

14. Smith M. and Kanade T. Video skimming and characterization through the combination of image and language understanding techniques. Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR), 775-781. 1997. IEEE.

15. Stifelman, L. The Audio Notebook: Paper and Pen Interaction with Structured Speech *Ph.D. dissertation, MIT Media Laboratory*, 1997.

16. Stifelman, L.J., Arons, B., Schmandt, C. & Hulteen, E.A. VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker. *Proc. INTERCHI'93 (Amsterdam, 1993),* ACM.

17. Tonomura, Y. & Abe, S., Content Oriented Visual Interface Using Video Icons for Visual Database Systems. In *Journal of Visual Languages and Computing*, vol. 1, 1990. pp 183-198.

18. Zhang, H.J., Low, C.Y., Smoliar, S.W. and Wu, J.H. Video parsing, retrieval and browsing: an integrated and content-based solution. In *Proceedings of ACM Multimedia, September 1995*, pp. 15-24.