

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 summarization and skimming tools so that they can find the presentation they want and browse through it quickly. In a previous study, we reported three automated methods for generating audio-video summaries and a user evaluation of those methods. 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) an audio-video summary created by the speaker. Results show that although some text-only conditions can match the audio-video summary, users have a marginal preference for audio-video (ANOVA $f=3.067$, $p=0.087$). 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, video summarization, digital video library, video browsing, video skim, multimedia.

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. 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 summarization 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* [3,12]. It allows the *complete* audio-video to be watched in a shorter amount of time by speeding up the playback with no pitch distortion. The achievable speed-up is about 1.5-2.5 fold depending on speaker [3], beyond which the speech starts to become incomprehensible. Higher speed-ups are possible [5], but at cost of increased software complexity and listeners' concentration and stress level.

Getting a much higher time saving factor (2.5+) 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 slide transition in a 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 reading the highlighted text summary or watching the audio-video summary?

Motivated by these questions, we compare four conditions: slides-only, full text-transcript with no highlights, full text-transcript with highlights, and audio-video summary. We also compare the results of this study and our earlier one [8]. We find that although comprehension is no different for full text transcript with highlights condition and audio-video summary condition, users have a subjective marginal preference for audio-video summary (ANOVA $f=3.067$, $p=0.087$). 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: 1) *slide-transition* points based

(S); 2) *pitch activity* based (P), using a modified version of algorithm introduced by Arons [3]; 3) 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 transcript¹. Below we describe each of these algorithms in slightly more detail.

The slide-transition-based algorithm (S) uses the heuristics that slide transitions indicate change of topic, and that relative time spent by speaker on a slide indicates its relative importance. Given a target summarization ratio of N%, the algorithm selects the first N% of audio-video associated with each slide for the summary.

The pitch-based algorithm (P) is based on research indicating that pitch changes when people emphasize points. In presentations, the introduction of a new topic often corresponds with an increased relative pitch, and pitch is more discriminating than loudness, etc. The pitch-based summary uses a slight variation of the algorithm proposed by Arons [3] to identify emphasis in speech. The use of this algorithm allows us to compare the new algorithms with this seminal earlier work.

The third algorithm (SPU) combines the information sources used by the two algorithms above and adds information about previous users viewing patterns, in particular, the number of distinct users that had watched each second of the talk. We used the two heuristics: 1) If there is a surge in the viewer-count of a slide relative to the previous slide, it likely indicates a major topic transition. 2) If the viewer-count within a slide falls quickly over time, the slide is likely not interesting. These heuristics are used to determine a priority for the slide, and each slide in the talk gets a fraction of the total time based on its priority. Given a time quota for a slide, we use the “pitch-activity” based algorithm to pick the segments included in the summary.

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.

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

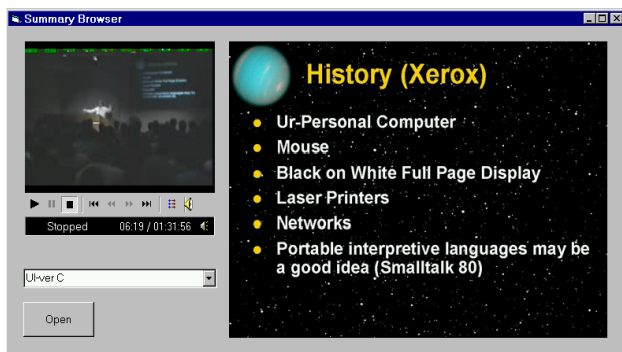


Figure 1: The interface for the experimental software.

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 for our evaluations: 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 (ANOVA $f=16.088$, $p=0.000$). The automated methods also resulted in substantial quiz score improvement, but the amount of improvement was statistically indistinguishable from each other (ANOVA $f=0.324$, $p=0.724$).

One hypothesis for lack of significant difference between the automated methods was that most of the useful information may be coming 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 infrequent). However, participants estimated that slides carried only 46% of the information and audio-video carried 54%. So the hypothesis is not quite justified; one goal of the current study is to help resolve this issue.

Survey responses also indicated a preference for author-generated summaries (see details later in Table 4). While along one dimension—the summary provides a good synopsis of the talk—the author-generated and automated summaries did comparably (ANOVA $f=0.521$, $p=0.472$), the author-generated summaries were preferred along other dimensions (all p 's were less than 0.005 using ANOVA),

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), and 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 so that a comparison with the results of previous study is possible.

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, majority of the time preparing for a talk 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 as a summary. 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 spoken language is informal, it contains filler words and repetitions, and it may often be grammatically incorrect. 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 points in the transcript highlighted. We had the option of having key points marked by human experts or use the ones generated by our automated algorithms. We chose to use the ideal

case, the transcript highlighted by an author or expert. The reasons were two fold. First, this choice allowed us to compare the effectiveness of exactly the same summary when delivered with and without audio-video. Second, we believed this choice increases the longevity of results of this paper, as the quality of automated summaries will keep evolving (making future comparisons difficult), while those generated by humans should be quite stable.

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 the results. The four talks were on user-interface design (UI), Dynamic HTML (DH), Internet Explorer 5.0 (IE), and Microsoft Transaction Server (MT), respectively.

Table 1: Information associated with each presentation.

	UI	DH	IE	MT
Duration (mm:ss)	71:59	40:32	47:01	71:03
# of slides in talk	17	18	27	52
# of slides / min	0.2	0.4	0.6	0.7
# of words in transcript	15229	8081	6760	11578
# of pages in transcript	15	10	8	15
Highlighted words	19%	24%	25%	20%
Duration 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. The raw transcription texts are quite voluminous -- between 8 to 15 pages. They are hard to read even with the paragraph breaks on the slide transition points. In the summaries, the fraction of words highlighted by the speaker in the summaries is about 19 to 25%. Obviously, the end results may be different if much less or much higher summarization factors were chosen. A factor of 4 to 5 summarization seemed an interesting middle ground to us.

EXPERIMENTAL DESIGN

The same measures were used for outcome as for the first study: quizzes for 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 questions from different talks were jumbled up to reduce memorization of questions.

The 24 participants were employees and contingent staff members of Microsoft 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.

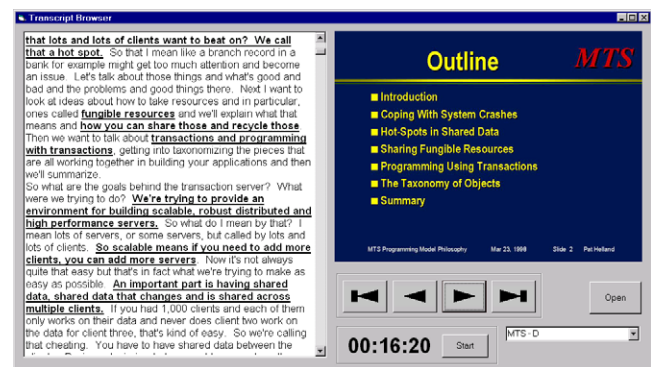


Figure 2: Interface for the conditions SO, T, and TH. The participant can use the vertical scrollbar to navigate the text transcript or use the four control buttons (shown below the slides) to navigate the slides. However, the current slide and the displayed text transcript are not linked. This allows 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:

1. Are there significant differences between the author-generated video summary and the text transcript with the same portion highlighted? What is the value of audio-video?
2. How much worse are SO and T compared to A and TH? This focuses on understanding the value of effort put in by authors in identifying highlights.
3. Are there performance differences across the talks? Is there something in the structure and/or organization of talks that affects performance?

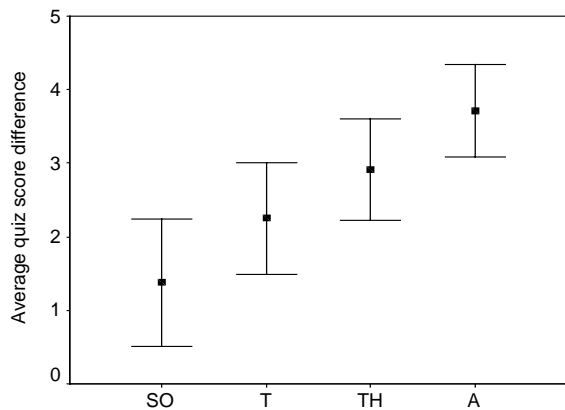


Figure 3: Average quiz score improvement by condition with 95% confidence intervals. The apparent linearity of the quiz score improvement is coincidental.

Figure 3 shows the average 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.

To more specifically answer the first question raised above, we compared A and TH conditions. Analysis of data shows that there is only a marginal preference for audio-video (ANOVA $f=3.067$, $p=0.087$). We also had subjective comments from users about added value from hearing the speaker’s voice and intonation. We present these comments in “User Comments” subsection below.

To answer the second question, we analyzed the data with quiz scores from conditions A and TH as one group, and those from T and SO as another group. Data show that A and TH are significantly better than SO and T as a group (ANOVA $f=16.829$, $p=0.000$). Thus, the author effort in

identifying highlights does add significant value to the viewers.

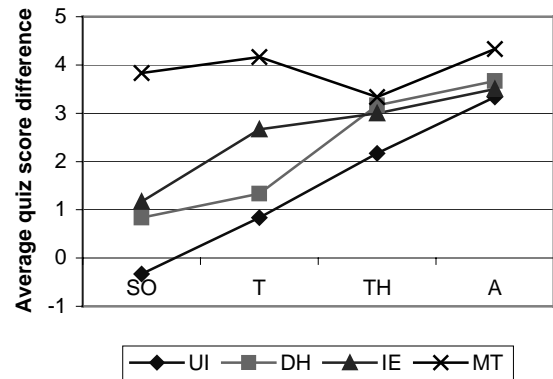


Figure 4: Variation in quiz score for each talk as a function of summary method.

To answer the third question, we show quiz score improvement as a function of summary method for each of the talks in Figure 4. Two things stand out. First, the variation in score improvement across summary methods is the least for the MT talk and the most for the UI talk. Second, the variation in score improvement across talks is the least for summary method A and the most for summary method SO.

As we looked deeper, we found the amount of variability in the quiz score improvements across methods seemed to correlate with the extent of information present in the slides. For example, one measure of information density in slides 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). This corresponds to the observation that the variance as a function of summarization method is the least for MT talk and the most for UI talk.

Our intuition regarding this is that slide content is a key source of information for the viewers. If there is lots of detail in the slides (e.g., MT has 52 slides) then the summary method matters much less than if there is little information in the slides (e.g., UI has 17 slides). Conversely, a good summary can compensate for lack of information in the slides by suitably providing audio-video segments where key points are made.

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. The goal of the surveys was to get subjective reaction of users to the summary methods.

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

User Ratings

The pattern of responses was similar to that of the quiz scores (see Table 2 and Table 3). Average ratings for the video summaries (A) tended to be the highest. However, none of the seven ratings in Table 2 were significantly different between the audio-video summary and the highlighted transcript (TH) using ANOVA at $p=0.05$. For A and TH as a group, all of the ratings were significantly greater than those for the slides-only (SO) and transcript (T) using ANOVA at $p=0.01$. SO was significantly worse than the other three on all ratings, using ANOVA at $p=0.05$, except Enjoy (ANOVA $f=2.131$, $p=0.148$).

Table 2: Post-quiz survey results by conditions³. Responses were from 1 (“strongly disagree”) to 7 (“strongly agree”).

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

Also following the quiz score trend is the fact that the average ratings for the MT talk were consistently higher than the others (see Table 3), independent of the summary method. Again this is likely due to the fact that the slides were sufficiently detailed so that they could “stand alone” and be interpreted without the speaker’s audio-video/text-transcript.

Table 3: Post-quiz survey results by talks. Responses were from 1 (“strongly disagree”) to 7 (“strongly agree”).

By talk	Synop.	Effic.	Enjoy	Key points (%)	Skip talk	Concise	Coher.
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.

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.”

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 allowed them to listen passively, it was 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 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: 1) The talks and quiz questions were the same; 2) The author-generated audio-video summary (condition A) was present in both studies; 3) Slides were shown in all conditions in both studies; and 4) The method used to evaluate outcome was essentially the same. These similarities allow us to compare the results from these two studies.

Figure 5 shows the average quiz score difference by conditions for each talk 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. Also, audio-video is present in all conditions in the old study, compensating for level-of-detail differences in the slides.

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 we see its ratings are consistently lower in the present study. One hypothesis is that the ratings are relative to the quality of other conditions explored in the same study. The closeness of TH to A in current study might have resulted in A getting these slightly lower scores.

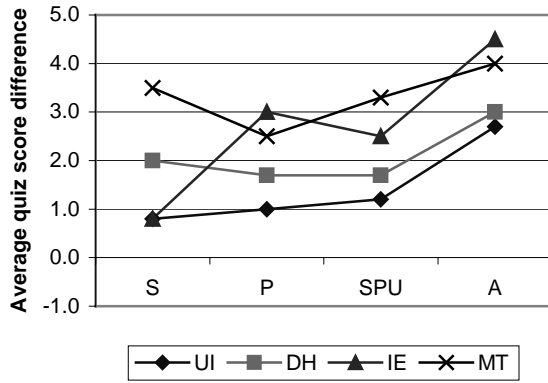


Figure 5: Average quiz score difference by conditions from the automatic summary study.

It is interesting to compare the S and SO conditions across the two studies. 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 SO⁴, 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
Old Study	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
Current Study	A	4.96	68.91	4.41	5.13	4.13
	TH	4.70	64.13	4.52	4.52	4.35
	T	3.58	61.67	3.83	3.50	4.17
	SO	3.13	41.25	1.96	2.92	2.83

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 (ANOVA $p=0.048$), less choppy (ANOVA $p=0.001$), and of higher quality (ANOVA $p=0.013$) than were the first three

⁴ This is true even if we adjust the ratings of S using $S' = S * A / A'$ because of the difference between A and A' in the two studies.

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.

RELATED WORK

There has been considerable research on indexing and searching the rapidly expanding sources of digital videos [1,2,5,10,11,13,15,18,19]. All these systems use automatic techniques based on the visual aspect of the media, primarily employing image-recognition and image-processing techniques (e.g., shot boundaries). Some of them [11,15] use textual information from speech-to-text software or closed captions. While these systems focus on the technical aspect, our study focuses on the human side: comparing the effectiveness of different summary abstractions for audio-video presentations.

Christel et al. [4] report a subjective evaluation of video summaries created from image analysis, keyword speech recognition, and combinations, for general-purpose videos. Based on analysis, summaries (or what they call skims) are constructed by concatenating 3-5 second video segments. 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. This paper, in contrast, focuses on presentations allowing domain specific knowledge to be used (e.g., 3-5 second speech segments used by Christel et al. are too short for comprehension).

The interfaces we used in the user study were simple (see Figure 1). One can imagine more sophisticated interfaces that takes the advantage of the digital media. Barry Arons' SpeechSkimmer [3] allows audio to be played at normal speed, or speeded-up with no pitch distortion, with pauses removed, or restricted to phrases emphasized by the speaker. Lisa Stifelman introduced Audio Notebook [16,17], a prototype note-pad combining pen-and-paper and audio recording. Audio Notebook allowed the user to use notes made on paper (ink marks) as an index into the recorded audio presentation. In contrast to SpeechSkimmer and Audio Notebook that focused on audio alone, Li et al. [9] explored interfaces addressing both audio *and* video, providing enhanced features such as time-compression, pause removal, navigation using shot boundaries and table-of-content. Integrating these enhanced browsing features with the summary methods studied here could substantially enrich the end user's experience.

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 summarization 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.

We show that slides-only (S) and plain transcript (T) summary methods are significantly worse than author generated transcripts with highlights (TH) and audio-video (A) summary methods. Author participation clearly adds value. We also show that while comprehension given transcripts with highlights method (TH) can match the audio-video summary (A), there is a marginal preference for audio-video (ANOVA $f=3.067$, $p=0.087$). Furthermore, we observe that different styles of slide-authoring (e.g., detailed vs. major points only) can have a large impact on their effectiveness as summaries. The results conflict with the common advice that slides should contain only the major points to retain attention of live audience. This raises a dilemma for speakers who are authoring for both on-demand and live audiences. One 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 Multimedia'96*, pp 1-10. ACM.
2. Arman, F., Depommier, R., Hsu, A. & Chiu M.Y. Content-based Browsing of Video Sequences, In *Proceedings of Multimedia'94*, pp 97-103. ACM.
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 Browser using Automatic Multimodal Analysis. In *Proceedings of Multimedia'98*, pp. 375-380. ACM.
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. Li, F.C., Gupta, A., Sanocki, E., He, L. & Rui, Y., 1999. Browsing Digital Video. In *Proc. CHI 2000*. ACM.
10. Lienhart, R., Pfeiffer, S., Fischer S. & Effelsberg, W. Video Abstracting, *ACM Communications*, December 1997.
11. Merlino, A., Morey, D. & Maybury, M. Broadcast News Navigation Using Story Segmentation. In *Proceedings of the 6th ACM international conference on Multimedia*, 1997.
12. 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.
13. 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.
14. Stanford Online: Masters in Electrical Engineering, 1998. <http://scpd.stanford.edu/cee/telecom/onlinedegree.html>
15. 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*, 775-781. 1997.
16. Stifelman, L. The Audio Notebook: Paper and Pen Interaction with Structured Speech *Ph.D. dissertation, MIT Media Laboratory*, 1997.
17. 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.
18. 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.
19. 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.

