Auto-Summarization of Audio-Video Presentations

Liwei He, Elizabeth Sanocki, Anoop Gupta, and Jonathan Grudin
Microsoft Research
Redmond, WA 98052

ABSTRACT

As streaming audio-video technology becomes widespread, there is a dramatic increase in the amount of multimedia content available on the net. Users face a new challenge: How to examine large amounts of multimedia content quickly. One technique that can enable quick overview of multimedia is video summaries; that is, a shorter version assembled by picking important segments from the original.

We evaluate three techniques for automatic creation of summaries for online audio-video presentations. These techniques exploit information in the audio signal (e.g., pitch and pause information), knowledge of slide transition points in the presentation, and information about access patterns of previous users. We report a user study that compares automatically generated summaries that are 20%-25% the length of full presentations to author generated summaries. Users learn from the computer-generated summaries, although less than from authors' summaries. They initially find computer-generated summaries less coherent, but quickly grow accustomed to them.

Keywords: Video summarization, video on-demand, streaming media, corporate training, digital library, user log analysis, user evaluation

1 INTRODUCTION

Digital multimedia content is becoming pervasive both on corporate Intranets and on the Internet. For example, many corporations are making audio and video of internal seminars available online for both live and on-demand viewing—at Microsoft Research about 5-10 internal seminars are made available online every week. Similarly, many academic institutions are making lecture videos and seminars available online. For example, Stanford University is currently making lecture videos from ~25 courses available online every quarter (http://stanfordonline.stanford.edu), and research seminars from Stanford, Xerox PARC, University of Washington, and other sites can be watched at the MURL Seminar Site (http://murl.microsoft.com). 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.

We humans are great at skimming text to discover relevance and key points; in contrast, multimedia content poses challenges because of the temporal rather than spatial distribution of content. In this paper we focus on techniques for automatic creation of summaries for audio-video presentations; that is, informational talks given with a set of slides. Auto-summarization is particularly important for such informational content as contrasted to entertainment content, such as movies. Entertainment content is more likely to be watched in a leisurely manner and costs so much to produce that it is reasonable to have a human produce previews and summaries.

The techniques we propose for summarizing informational talks exploit information available in the audio signal (e.g., pitch and pause information), knowledge about slide transition points in the presentation, and information about access patterns of previous users. We present results from user studies that compare these automatically generated summaries to author generated summaries.

Several techniques have been proposed for summarizing and skimming multimedia content to enable quick browsing. In one class of schemes [1,2,18,25,26], a static storyboard of thumbnail images is built from the video channel using information such as amount of motion or the newness of visual context. These techniques are less applicable to informational presentations, in which the video channel consists primarily of a talking head and most of the useful information is in the audio channel.

Another class of techniques is that of time compression [3,5,17,23,24]. These can allow the *complete* audio-video to be watched in a shorter amount of time by speeding up the playback with almost no pitch distortion. These techniques, however, allow a maximum time saving of a factor of 1.5-2.5 depending on the speech speed [10], beyond which the speech becomes incomprehensible. Removing and shortening pauses in speech can provide another 15-20% time reduction [9].

In this paper we explore techniques that allow greater timesavings by removing portions of the content rather than just playing it faster (although time-compression can be added to provide further savings). The result is a "summary" that highlights the content and can be 4 to 5

times shorter. Previous techniques in this area have varied from straightforward sampling of the time line (e.g., play 10 seconds, skip 50 seconds, play 10 seconds, etc.) to more sophisticated techniques based on pitch, pause, speech energy, speech-to-text content analysis, and video channel analysis [1,2,7,8,12,16,22].

We present a detailed comparison with related works in Section 7, but at a high level, the work reported here differs as follows: First, we focus specifically on informational presentations with slides. This context provides us with a critical new source of information, the time-points when slides are changed, allowing new algorithms. Second, we use the access patterns of users who have watched the talk prior to the current user. Finally, we present detailed comparison with author-generated summaries, something that most of the earlier studies do not do.

The paper is organized as follows: Section 2 describes our study context and discusses sources of information available for the summarization task. Section 3 presents the three automatic video summarization algorithms proposed in this paper. Section 4 gives details of the experimental design of our user study. Sections 5 and 6 present the results. We discuss related work in Section 7 and conclude in Section 8.

2 SOURCES OF INFORMATION

As stated earlier, our focus is on creating summaries for informational talks. Let us begin by enumerating the characteristics of this domain and the sources of information available to it.

2.1 Informational Talks

Informational talks are often accompanied by slides. As concrete example, we consider seminars made available online by the Microsoft Technical Education (MSTE) group.

The displays of online courses are usually very similar. Figure 1 shows the typical view for a MSTE seminar: a *video-frame* with VCR controls (play/pause/fast-forward/rewind), a *table-of-contents frame* (lower left corner) where users can click to go to that particular section of the talk, and a *slide-frame* where the users can see the slide relevant to the current audio-video content. As the video progresses, the slide-flips are automatic, requiring no user intervention. The slide-frame is often needed, because the video window is too small to show the detailed content of the slides.

While the process of authoring such content used to be quite cumbersome, it is quite simple nowadays. With the deployment of Microsoft PowerPoint Presenter, the timing of each slide-switch is recorded in real-time as the speaker flips slides on the laptop, and linked in with the audio-video stream being recorded at the video server.

The client-server architecture also makes user action logging possible. Every time a user interacts with a control element, such as an item in table-of-contents (TOC) or a video speed control, an event is generated. The interface can be instrumented so that each event is time stamped and logged, along with the user's ID, as a database record on the server.

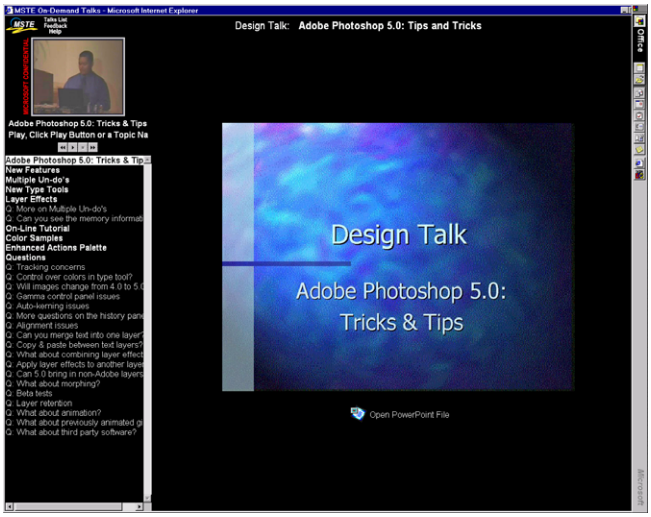

Figure 1: User Interface for MSTE talks.

2.2 Sources of Information for Summarization

At a high level, we can consider four major sources of information that can be used in automatically producing summaries: 1) video channel; 2) audio channel; 3) speaker's actions; and 4) end-users' actions watching the talk. Unfortunately, for the talks considered here, the video is primarily a low-resolution face shot. Although we might get some useful information from the degree of animation of the speaker and hand gestures that are captured, for this study we ignore this channel and focus on the other three sources of information.

2.2.1 Audio channel

In informational talks, audio carries much of the content of the presentation. There are two main sources of information here: 1) the spoken text or the natural-language content, and 2) the pitch, pause, intensity, intonation, and other prosody information in audio. In this paper, we focus on the latter.

Research in speech and linguistics communities [11,13,20] has shown that changes in pitch occur under various speaking conditions. The introduction of a new topic often corresponds with an increased pitch range. Pitch activity provides information in speech that is important for comprehension and understanding. In this paper, we explore the use of pitch to identify the speaker's emphasis in the presentation. Our algorithm is based on Barry Aron's work [4]. We use it both by itself (providing comparison to Barry's earlier work), and in combination with new sources of information discussed below. We also use pause information to avoid selecting segments that start in the middle of a phrase, which can be annoying.

2.2.2 Speaker actions

In addition to a speaker's audio, there is useful information in the speaker's actions: gestures, when he/she transitions to

the next slide, how much time he/she spends on a slide, and so forth. As stated earlier, for this study we ignore gesture information in the video channel, but we do consider the information in slide transition points.

This information is easily and automatically gathered and contains two potentially important elements: 1) A slide transition usually indicates the introduction of a new topic or sub-topic, revealing the logical structure of the presentation; 2) The duration between slide transitions, the time spent on each slide, provides an indication of a slide's relative importance.

2.2.3 End-user actions

It is unlikely that a user going to watch an online talk is the first person to do so. Many people have likely watched that seminar before him/her, and we suggest using this as an additional source of information for creating summaries.

On the MSTE video server, for research purposes, information is logged regarding the segments of video watched by each user. We have 117,000 records from 6685 viewing sessions for 163 different presentations. Figure 3 shows the data accumulated across all sessions that had watched this given talk for over 10 minutes. The dotted vertical lines denote the slide transition points.

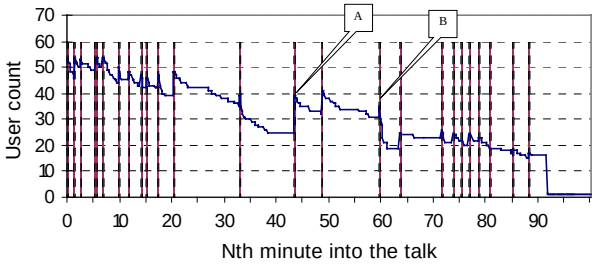

Figure 2: User-count vs. time plotted for one of the MSTE presentations. The vertical lines denote the slide-transition points. Note the spikes in user counts at the beginning of each slide transition point indicating users' jumping using table of contents.

There are several trends that we can see. First, in general, the number of viewers decreases as the time progresses. This is understandable as people start at the beginning and given limited time, opt out at some time before the end. Second, we see jumps in user-count at the beginning of slides. This is due to the availability of table-of-contents with the talk. When users get bored with the content of the slide, or think they have gotten the essence, they may decide to jump to the next interesting slide. This also accounts for the observation that within a slide, the number of viewers decreases as time progresses.

Thirdly, in contrast to the first observation, we see that for some slides (e.g., marked by label "A" in Figure 3) the average number of users increases as compared to the previous slide. We found that this corresponded to major topic transitions in the talk, and hence was significant information usable in the summary.

Finally, we see that for some slides, the number of users that watched the audio associated with that slide dropped very rapidly after the beginning of the slide (e.g., marked by label "B" in Figure 3). This indicates while the title of slide might have been interesting, its contents were most likely not interesting — gain a useful hint for us in the summarization task. We present how we use all this information in the next section.

2.3 Desirable Attributes of a Summary

Before going into summarization algorithms, it is good to think about the desirable attributes for an ideal summary. We believe the following four properties, the 4 Cs, offer a starting point.

1. **Conciseness**: Any segment of the talk that is selected for the summary should contain only necessary information. Thus, each segment should be as short as possible.
2. **Coverage**: The set of segments selected for the summary should cover *all* the "key" points of the talk.
3. **Context**: The segments selected and their sequencing should be such that prior segments establish appropriate context (e.g., critical terms should be defined before being used in a segment).
4. **Coherence**: The flow between the segments in the summary should be natural and fluid.

Of course, even an average human editor won't consistently produce summaries that satisfy all four criteria. However, we asked our users to evaluate the summaries along these dimensions.

3 THE SUMMARIZATION ALGORITHMS

We developed three automatic summarization algorithms. The first algorithm (denoted by S) uses information in *slide-transition* points only. The second algorithm (denoted by P), introduced by Arons [4], tries to identify emphasized regions in speech by *pitch activity* analysis. The third algorithm (denoted by SPU) uses all three sources of information: the slide-transitions, pitch activity, and user-access patterns.

In our pilot studies, users found it very annoying when audio segments in the summary began in the middle of a phrase. For all algorithms, we use a heuristic to ensure this does not happen by aligning segments to pauses in speech. The pauses in speech are detected using a pause detection algorithm proposed by Gan [9].

3.1 Slide-transition Based Summary (S)

The slide-transition-based summary 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 summary duration, the algorithm therefore allocates time to each slide in proportion to what the speaker did in the full-length talk. Furthermore, we use the heuristic that important information is spoken at the beginning of a slide, so we take

the appropriate duration segment right after each slide transition and add it to the summary.

3.2 Pitch-activity Based Summary (P)

The pitch-based summary uses the algorithm proposed by Arons [4] to identify emphasis in speech. The use of this algorithm allows us to compare the new algorithms with this seminal earlier work.

The pitch tracker we use is loosely based on an algorithm developed by Medan, Yair, and Chazan [15]. Another key attribute is the length of individual audio segments that are selected for the summary. In contrast to Arons, who had chosen 8-second segments, we decided to target roughly 15-second segments. This was based on the segment lengths we found in author-generated summaries (see Table 3). The algorithm works as follows.

1. The audio wave file is divided into 1ms frames.
2. The pitch tracker computes fundamental pitch frequency for each frame.
3. A threshold is chosen to select the pitch frames containing the top 1% of pitch values.
4. The number of frames in each one-second window that are above this threshold is counted to provide a measure of "pitch activity."
5. A sliding-window algorithm is used to determine the combined "pitch activity" for each 15-second window. They provide a measure of emphasis for phrase- or sentence-sized segments of a speech recording.
6. The 15-second windows are sorted according to the combined score.
7. Segments in the windows are added to the summary in the order of decreasing combined score until the summary reaches its specified length.

3.3 Summary Based on Slide, Pitch, and User-Access Information (SPU)

This algorithm combines the information sources used by the two algorithms above and adds information from user-access patterns. In developing the algorithm, a key question was how to incorporate this user-access information.

Our first attempt was to use the information fairly directly, that is, if a lot of users watched a given 10-second segment of the talk, then we would give that segment extra weight to be included in the final summary. We finally decided against this straightforward approach because of the biases in our data collection approach. First, most users tend to listen to a talk from beginning to end until they run out of time or are interrupted and have to leave, etc. Thus, there is unnecessary emphasis given to the front part of the talk (see Figure 3), even though there may be important content later on. Our data showed that for high summarization factors, the latter part of the talk got totally dropped. Second, the table-of-contents markers can only take users to the beginnings of the slides. So similar to first point, at a finer granularity this time, there is unnecessary emphasis given to the content right after a slide transition (see Figure 3).

Given the above biases, we decided to exploit user-access information in a fairly coarse-grained manner for this study – to prioritize the relative importance of slides. In future, we plan to study its use in finer-grained ways. To do this, we used the latter two heuristics mentioned in Section 2.2.3: 1) If there is a surge in the user-count of a slide relative to the previous slide, it likely indicates a topic transition. 2) If the user-count for a slide falls quickly, the slide is likely not interesting. The specific quantitative measure we used in this paper was the ratio between the average-user-count of the current slide and the average-user-count of the previous slide. Here is an outline of the algorithm we used.

1. Compute the importance measure for each slide as the ratio average-user-count for current slide divided by that for previous slide. Since there is no previous slide for first slide, it is always marked as the most important.
2. The slides in the talk are partitioned into three equal-sized groups based on the metric computed above: important, medium, and unimportant. The slides in the first group are allocated 2/3 of the total summary time; the medium group slides are allocated 1/3 of the total summary time; the least important slides are allocated no summary time. (1/3 and 2/3 were selected heuristically.)
3. Within each group, each slide gets a fraction of the total time allocated to that group using the slide-transition based algorithm described in Section 3.1. (This is how we incorporate slide-transition information into the algorithm.)
4. Given a time quota for a slide, we use the "pitch-activity" based algorithm outlined in Section 3.2 to pick the segments included in the summary.

3.4 Author Generated Summaries (A)

To establish a reference point for the effectiveness of summaries created using auto-summarization techniques, we asked human experts to produce summaries of the talks. Experts were defined as the author of the talk or someone regarded as qualified by that author.

Each author was given a text transcript of the talk divided into sections based on the places where slide transitions occurred. Authors marked summary segments on the transcripts with a highlighting pen. These sections were then assembled into a video summary by aligning the highlighted sentences with the corresponding video.

For the subsequent user study, each author also wrote 9 to 15 quiz questions that covered the content of their talk summary. Questions required participants to draw inferences from the content of the summary or to relay factual information contained in the summary. Either way, the quizzes were used to determine whether the computer-generated summaries had captured the content of the talk deemed relevant by the author.

4 EXPERIMENTAL METHOD

Evaluating summarization algorithms is a fundamentally difficult task. For example, the criteria we listed in Section 2.3 are both highly complex and subjective. As a result, the primary method we employ is a user-study with human subjects. We begin by describing the talks used in the study.

4.1 Talks Used in the Study

For our study, we chose four presentations from the MSTE internal web site. The criteria used were:

- The presentations had been viewed in the past by 30+ users to ensure that the usage logs were meaningful.
- Each presentation covered one topic rather than comprising a number of autonomous sections.
- An expert (preferably the author of the talk) was available to create the summary and quiz questions.

The four talks we chose were on the topics of user-interface design (UI), Internet Explorer 5.0 (IE), Dynamic HTML (DH), and Microsoft Transaction Server (MT). Table 1 below shows the number of slides used by the author and the original duration, and the number of users for each of the talks.

Table 1: Number of slides, length (mm:ss), and the user count of the chosen talks.

	UI	IE	DH	MT
Slide count	17	27	18	52
Original Length	71:59	47:01	40:32	71:03
User count	68	37	41	38

4.2 User Study Design

To compare the summarization techniques we used a combination of two methods. 1) We administered quizzes to users before and after watching summaries to quantify the effectiveness of summaries in capturing key content. 2) We used surveys to gauge participant's subjective reactions to the summaries.

We used 24 subjects in our study. This number allowed us to construct a procedure where we could completely counterbalance against ordering effects, i.e., the order in which users experienced the summarization methods and the talks.

The 24 participants were internal Microsoft employees and contingent staff members working in technical job positions. All lacked expertise in the topic areas of the talks. No mention was made until study sessions were over that computer algorithms were used to construct the summaries. Employees were rewarded with a coupon for a free espresso drink; contingent staff members were given the option of an espresso drink coupon or a free software product.

Participants first completed a background survey and took a 32-question multiple-choice pre-study quiz to document their expertise in the topic areas of the talks. Note, these were simply the combined questions generated by all four authors of the talks. We took the precaution of jumbling up the ordering of questions within and across talks, so that the subjects had less recollection of the questions while watching the talk summaries and at post-summary quiz time.

Each participant watched four summaries, deploying a different summarization technique across the four talks. The order in which talks and summarization techniques were presented to each subject was such that we counterbalanced against all ordering effects.

While watching a summary, a participant could pause and jump back and jump forward. Once finished watching, however, participants were instructed to stop and not review portions of the summary. Participants were provided pen and paper to take notes. Figure 3 shows the user interface seen by the subjects.

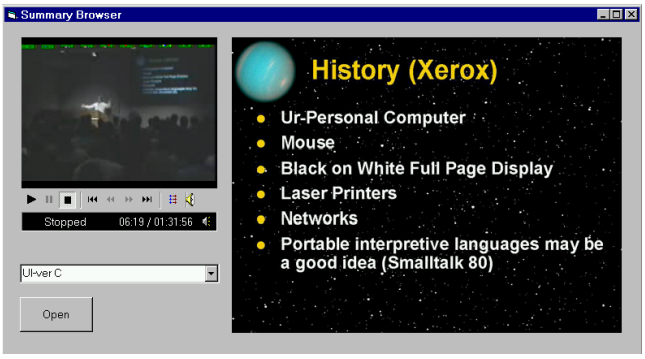

Figure 3: User interface for the subject testing software. The slides, shown on the right, are synchronized with summary-segment transitions.

After watching each summary, participants filled out an opinion survey and took a post-summary quiz. The survey evaluated the participants' qualitative experience watching the summaries. The score difference between the pre-summary and the post-summary quiz measured the amount of new information gained.

5 CHARACTERISTICS OF SUMMARIES

Before discussing the results of the user-study, we briefly comment on some basic quantitative characteristics of the summaries produced by the four methods.

Table 2: Duration of original talk and summaries (mm:ss).

	UI	IE	DH	MT
Original	71:59	47:01	40:32	71:03
A	13:44	11:37	9:59	14:20
SPU	13:48	12:08	9:54	14:16
P	14:14	12:06	9:58	14:03
S	13:58	11:36	10:19	14:30
Summar. Factor	**5.2**	**4.0**	**4.1**	**5.0**

Table 2 lists the duration of the original presentations and the four summaries, and the summarization factor. There is some variation on how much the different talks are summarized. We had given the authors a target of roughly $1/5^{th}$ the duration, and we naturally got some variation in how much of the text transcript they marked up. Once we got the author summaries, we gave the auto-summarization algorithms the same target duration. Again, we see some variation in the duration between different methods for the

talk. This is because the segments are extended to the next phrase boundary (see Section 3).

Table 3: Average length and standard deviation (in parentheses) of the summary segments in seconds.

	UI	IE	DH	MT
A	15 (12)	21 (15)	19 (15)	15 (10)
SPU	19 (4)	20 (7)	18 (3)	18 (6)
P	18 (2)	18 (3)	18 (2)	19 (7)
S	48 (43)	25 (22)	33 (25)	16 (8)

Another interesting attribute that we can examine is statistics about the length of individual segments constituting the summaries. In the past, Arons [4], Christel [7] and Smith and Kanade [22] have chosen relatively small segments of 2-8 seconds for their browsing/highlight generation applications. Before conducting the study, we were curious what the author-generated summaries would look like.

Table 3 shows average segment length and standard deviation across the summarization algorithms and talks while Figure 4 shows detailed distributions averaged across the four talks. We see that, in contrast to earlier work, author-generated summaries have much longer segment lengths, with averages between 15-21 seconds and standard deviations of 10-15 seconds. To some extent this may be explained by the fact that when browsing, one simply wants better coverage with lots of shorter segments—we may expect the human to use other VCR-like controls to explore the surrounding region if he/she feels so inclined.

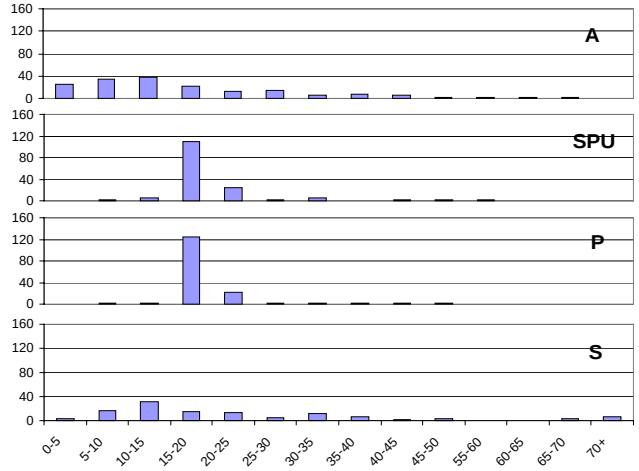

Figure 4: Distributions of segment lengths for different summarization algorithms. The horizonal axis contains bins of different segment lengths (in seconds). The vertical axis is the number of segments in a specific bin.

As for auto-summarization algorithms, based on our observations from the author-generated summaries, we tuned the P and SPU algorithms to have averages close to 15 seconds (see Sections 3.2 and 3.3). The variations that we see are present due to forced alignment with phrase boundaries identified by long pauses. For the slide-based algorithm (S), the average segment length and the variation are very large. The reason is that only the beginning portion of each slide is selected by the algorithm, and time that authors spend on each slide can vary a lot.

Another interesting metric to look at is if there is much overlap between the segments chosen by the various algorithms. This data are shown in Table 4 below. On one hand, if the emphasis detection algorithms were doing a great job, we should see more than just chance overlap between the author-generated and the auto-summarization algorithms. (The chance overlap percentages for the four talks are roughly 19% for UI, 25% for IE, 25% for DH, and 20% for MT.) On the other hand, given that author summarization leverages semantic understanding of the content while the auto-summarization algorithms don't, our expectations should be fairly low.

Table 4: Percentage of overlap between two different summaries within the talk.

UI	A	SPU	P	S	**DH**	A	SPU	P	S
A	100%				A	100%			
SPU	12%	100%			SPU	23%	100%		
P	15%	56%	100%		P	24%	67%	100%	
S	24%	18%	12%	100%	S	27%	23%	27%	100%
IE	A	SPU	P	S	**MT**	A	SPU	P	S
A	100%				A	100%			
SPU	25%	100%			SPU	21%	100%		
P	38%	54%	100%		P	15%	57%	100%	
S	36%	34%	33%	100%	S	25%	31%	39%	100%

We indeed see that there is not a significant overlap between the author-generated segments and the automatically generated segments (look down column A in Table 4). In fact, the overlap is the most with the slide-based algorithm, which covers the beginning content of each slide, indicating that that is a resonable heuristic. There is considerable overlap between the P and the SPU algorithms, as they both use pitch-activity as the primary emphasis detection algorithm, but that does not translate to significant overlap with the author-based summary.

Table 5: Slide coverage for the different summaries.

	UI	IE	DH	MT
Slide count	17	27	18	52
A	94%	67%	78%	50%
SPU	59%	78%	67%	77%
P	76%	85%	61%	71%
S	100%	100%	100%	100%

Finally, in the user-interface provided to our subjects (see Figure 3), the slides carry a substantial fraction of the useful information. So, we wanted to explore what fraction of the slides was shown to the subjects under the various algorithms. This data are shown in Table 5. For the author-

generated summaries we can clearly see that the fraction of slides exposed in the summary goes down as the total number of slides in the talk increases. This is not the case for the auto-generated summaries, which may be good (more information) or bad (too much clutter, and only a short exposure). We see users' opinions in the section below.

6 USER-STUDY RESULTS

Evaluating summarization algorithms is a fundamentally difficult task, as the critical attributes are highly complex and difficult to quantify computationally (e.g., unlike mean-square-error for visual distortion). As discussed in Section 4.2, we use a combination of performance on a quiz and ratings on an opinion survey for our evaluation.

6.1 Quiz Results

The purpose of the quiz was to see to what extent concepts that are considered "key" by the authors are captured and communicated by the various summarization methods. It was expected that the quizzes would produce best results for the author-generated summaries, since the quizzes were created by the authors to cover the material they selected. However, we wanted to know: 1) How close do automated summarization methods get to author-generated summaries? 2) Are there significant differences between the various automated summarization methods?

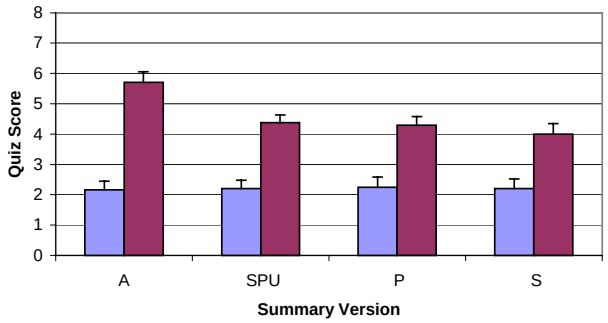

Figure 5: Pre-study (left bar) and post-summary (right bar) quiz scores for each summary type. The figure also shows the Standard Error for each bar. Improvement in scores for author-generated summaries was significantly higher than that for other schemes.

Figure 5 shows the absolute pre-summary and post-summary quiz scores obtained by the subjects as a function of the summarization method. As expected, we see essentially no differences in the pre-summary quiz results. For A, SPU, P, S methods the scores averaged 2.17, 2.21, 2.25, and 2.21 respectively out of a maximum of 8. After watching the summaries, the scores went up to 5.71, 4.38, 4.29, and 4.00 respectively. As expected, author-based summaries did better than the automated methods (statistically, the difference is significant at the 0.01 level.) However, in an absolute sense, the automated methods do not do too poorly: All post summary quiz scores increased significantly. Had different experts generated summaries and created quizzes, the quiz score differences would likely have been smaller.

As for our second question, there was no statistically significant difference among the automated methods. On the positive side, we may not need to bother with the more sophisticated methods, such as SPU, and rely on simpler methods such as S and P. On the negative side, it may simply indicate that we are not as yet exploiting user information effectively. The coarse-grain strategy described in Section 3.3 may not be paying off.

Another possible hypothesis for no difference may be that the key distinguishing feature between the algorithms is which audio-video segments are selected, but most of the useful information may come from the slides (which are shown independent of the specific audio-video portion played from that particular slide). We asked the subjects about the utility of slides explicitly in the opinion survey, so we comment about it in the next subsection. However, this hypothesis cannot be entirely true because the "S" algorithm alone displays 100% of slides to subjects (see Table 5), yet it does no better in subject performance.

6.2 Survey Results

Participants completed short preference surveys after watching each summary. These were conducted prior to taking the quiz so that their quiz performance would not affect their opinions.

The pattern of responses was similar to that of the quiz scores. The author-generated talk summaries were rated significantly more favorably than the computer-generated summaries. The data are shown in Table 6.

Table 6: User responses to quality of summary for various methods. The exact wording was: 1) Clear: "I found the summary clear and easy to understand." 2) Concise: "I feel that the summary was concise–it captured the essence of the talk without using too many sentences." 3) Coherent: "I feel that the summary was coherent–it provided reasonable context, transitions, and sentence flow so that the points of the talk were understandable." 4) Key Points: "My confidence that the summary covered the key points of the talk is:" 5) Skip talk: "I feel that I could skip the full-length talk because I watched this summary." Responses were on a seven-point scale (1 to 7) where 1 was "strongly disagree" and 7 was "strongly agree".

	Clear	Concise	Coherent	Overall quality	Key points (%)	Skip Talk
A	5.39*	5.57*	5.30*	5.09*	75.22*	4.87*
SPU	4.30	4.52	3.48	4.13†	63.04	3.43
P	4.00	4.13	3.48	4.09	63.04	3.05
S	4.26	4.17	3.64	3.55	58.26	3.30

* The mean difference is significant at the 0.05 level.

† The difference between the quality of the summaries for A and SPU was not significant at the usual .05 level but was significant at probability p= 0.09.

Considering specifically clarity, conciseness, and coherence, the computer-based methods were rated particularly poorly on coherence (down from ~5.3 to 3.5). Subjects complained that the summaries "jumped topics" and were "frustrating because it would cut out or start in the middle [of a topic]," and so forth. Although we attempted to

avoid cutting in the middle of phrases, a more elaborate technique may be required.

As for overall quality, as expected, Table 6 shows that the subjects preferred author generated summaries. However, many subjects were very pleasantly surprised to learn that three of the methods were computer based. Quotes included "Yeah, interesting. Wow! Then it's very cool." and "hum, that's pretty good for a computer. I thought someone had sat down and made them.." Real users will presumably be aware of the origins of summaries.

We also asked subjects about their confidence level that the summary had covered key points in the talk. Author-based summaries got a confidence-level of 75%; computer-based summaries got a respectable ~60%. We were expecting that the slide-based scheme would get high points here, because the subjects saw *all* the slides, but it got the lowest score. The fact that points in the middle and end of slides were not covered by the audio seemed to dominate judgments.

Finally, we asked subjects if they would skip the talk based on the summary. After hearing author-based summaries, subjects were more convinced that they could skip the talk.

Table 7: User responses to relative value of slides, audio, video in summaries. The exact questions was "What percent of the information do you feel came from the slides, the audio portion and the video portion of the summary (e.g. 30% slides, 30% audio, 40% video)?" Responses are in percent. Awareness asked the users to rate the extent "The video portion of the summary fostered a sense awareness of the speaker that I don't get from slides alone". Responses were on a seven-point scale (1...7) where 1 was "strongly disagree" and 7 was "strongly agree".

	Slides	Audio	Video	Awareness
A	47.17	38.26	14.57	5.04
SPU	44.32	35.14	18.09	5.04
P	42.61	39.13	18.26	5.00
S	50.65	30.22†	19.13	4.87

† The difference between the audio for S was significantly different for A and P, but not SPU.

Given our focus on informational presentations, we were interested in learning the value subjects felt they derived from slides vs. audio- and. video-tracks (Table 7). Clearly, most information was attributed to reading slide content (mean = 46.2%), followed by audio-track (mean = 35.7%) and video-track (mean = 18.1%). Clearly, it is imperative that slides be shown if at all feasible. Although the outcome is not particularly surprising in retrospect, it is useful to have this preference quantified.

There are no significant differences in the perceived value of slides, audio, and video across the summarization methods, except that the audio-track for author-based and pitch-based methods offers significantly more value than that for slide-based method (significant at the 0.05 level). We believe the reason is that the slide-based approach often has very short audio segments as it does not drop any slide, and these short segments often add no value at all to the subjects' understanding.

The contribution of the small video window found in many online presentation user-interfaces (see Figure 1, Figure 3) to the overall experience and/or sense of awareness of the speaker has been questioned. We asked this question of the subjects; the result is shown in the "awareness" column of Table 7. Given that 4.0 is the neutral value in this scale of 1 to 7, the score of ~5.0 indicates that subjects derive value from the video channel.

Table 8: Perception of clarity of summary (higher number better), choppiness of summary (lower number better), and overall quality of summary (higher number better) as a function of the order in which summaries were presented to subjects. Responses were on a seven-point scale where 1 was "strongly disagree" and 7 was "strongly agree".

Sequence in study	Clear	Choppy	Quality
1	4.04	6.00*	3.65
2	4.39	5.09	4.09
3	4.39	4.70	4.00
4	5.13*	3.91*	5.18*

* The mean difference is significant at the 0.05 level.

Finally, we observed significant habituation effects. Participants' perceptions of summary quality improved over time. Table 8 shows the results as a function of the sequence in which they watched a summary (the study was designed so that each of the four summary methods was presented equally often in each position in the sequence). The last summary shown in the study was consistently rated as being clearer (p=.048), less choppy (p=.001), and of higher quality (p=.013) than the first three summaries. This specific result lends a silver-lining to the generally very hard task of computer-based summarization: Over time, even with the early-stage summarization technology presented in this paper, users may indeed find considerable satisfaction from the summaries.

7 DISCUSSION AND RELATED WORK

There has been considerable research on indexing, searching and browsing the rapidly expanding sources of digital video [1,2,18,25,26]. These approaches all focus on visual aspects of media, primarily employing image-recognition and image-processing techniques.

This study focuses on informational talks accompanied by slides. It has several characteristics: 1) audio carries most of the information; 2) slides provide a logical structure to the talk; 3) user access information logged on the server could give further information as to which part of the talk is important. Since the video mainly consists of a talking head, storyboard or keyframe compression approaches used for other types of video are not effective.

Barry Arons' SpeechSkimmer [5] 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 [23,24].

Audio Notebook relies on the synchronization of the keypoints marked by pen on paper to provide structure to the recorded audio.

Our work shares some of the audio emphasis-detection framework with SpeechSkimmer and the Audio Notebook. However, it differs by focusing on informational talks, and thus exploiting information from sources such as slide transitions and end-user access logs recorded on a server. Both SpeechSkimmer and the Audio Notebook also focused on interactive browsing interfaces; in contrast, we focus on more passive summary viewing. Hence, the user-studies and the results are quite different. Our studies do reflect some on pitch-based emphasis detection algorithms proposed by Arons. For example, we see that the overlap between author-generated segments and pitch-based segments is not any more than what would be achieved by random chance.

Chen and Withgott [6] did a more comparable study in which an emphasis-recognition system was used to produce summaries of recorded telephone or interview conversations. The high correlation they found between emphasis and utility for inclusion in a summary suggests that emphasis may be used differently in brief, animated conversations and in long lectures. The study was also done by training a Hidden Markov Model on the first half of a recorded conversation and using it to predict emphasis in the latter half of the *same* conversation. It is not clear how the results will generalize to our situation, where no such training is possible.

More recently, in the CMU Informedia project, Smith and Kanade [22] produce a summary of non-lecture video automatically by image and language analysis. For example, if a face is recognized and a proper name appears on screen, they may conclude that a person is being introduced, and include a 2-second video shot in the summary. In contrast to our work, there is much greater emphasis on video channel, audio is used mainly for speech-to-text, and segments chosen are only 2 seconds long. They did not report user tests.

Also at CMU, Christel et al [7] report subjective evaluation of summaries created from image analysis, keyword speech recognition, and combinations, again from general-purpose video. Based on analysis, summaries/skims are constructed from video shots of 3-5 seconds each. They tested quality of skims using image recognition and text-phrase recognition tasks. Performance and subjective satisfaction of all skimming approaches contrasted unfavorably with viewing of the full video; the latter was negative for each technique on each dimension examined. Again, because of our study focus, we use different and new information sources not available to them. We also compare our techniques to author-generated summaries rather than watching full videos.

8 CONCLUDING REMARKS

With widespread availability of streaming-media solutions, ever decreasing storage costs, and increasing network bandwidths, we are seeing a dramatic increase in the amount of multimedia content available on the net. In the workplace and academic environments, a substantial fraction of such content takes the form of presentations consisting of audio, video and slides. In this paper we have presented techniques to automatically summarize such informational multimedia presentations.

Our focus on informational presentations allowed us to exploit new sources of information such as slide-transition timing and user-access logs that had not been explored before, in addition to traditional sources such as pitch-based emphasis detection. We described three algorithms based on these sources of information and compared them to author-generated summaries.

An analysis of author-generated summaries showed segment lengths of 15-21 seconds. This is in contrast to the 2-8 second segments generated by most earlier algorithms. We were also surprised to find little more than chance overlap between the segments generated by the proposed algorithms and the author-generated summaries. It appears presentations may be less susceptible to pitch-based emphasis analysis, or that spoken emphasis did not truly correspond to semantically important material.

User study with 24 subjects clearly showed that they preferred author-generated summaries to computer-generated ones. While the preference was large along some dimensions (e.g., they found author-generated summaries much more coherent), computer-generated summaries faired respectably along other dimensions (e.g., the confidence level that “key” points of talk had been covered by the summary).

Interestingly, we did not find any significant difference between users’ preferences for the three computer-based methods, leading us to conclude that the simpler methods (S and P) may be preferable for now. Overall, the computer-based summaries were well received by most subjects, and many expressed surprise when they learned that a computer had generated them. Another very surprising result was users increasing tolerance to computer-generated summaries. The last summary shown to the subjects was consistently rated as being clearer ($p=.048$), less choppy ($p=.001$), and of higher quality ($p=.013$) than the first three summaries, even though the study was designed so that each of the four summary methods was presented equally often in each position in the sequence. This indicates to us that the user community may be ready to accept current generation summarization methods, even though we know that there are many improvements to be made.

There are several future directions to be explored. We are conducting a study comparing audio-video summaries and text summaries of transcripts. We will be looking into exploiting other new sources of information (e.g., natural

language parsing and video analysis), fine tuning our use of current sources of information (e.g., more effective use of user-access logs as in collaborative-filtering systems [19]), understanding what role authors/speakers may play in helping generate better summaries (e.g., marking relative importance of slides), and making a more interactive and intelligent video browsing system, where the end-user will be a participant in a tight loop with the system.

ACKNOWLEDGEMENTS

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.

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. Techniques, Perception, and Applications of Time-Compressed Speech. In *Proceedings of 1992 Conference*, American Voice I/O Society, Sep. 1992, pp. 169-177.
4. Arons, B. Pitch-based Emphasis Detection for Segmenting Speech Recordings. In *Proceedings of International Conference on Spoken Language Processing*, vol. 4, 1994, pp 1931-1934.
5. Arons, B. SpeechSkimmer: A System for Interactively Skimming Recorded Speech. *ACM Transactions on Computer Human Interaction, 4*, 1, 1997, 3-38.
6. Chen, F.R. & Withgott M. The use of emphasis to automatically summarize a spoken discourse, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 229-233. 1992. IEEE.
7. 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.
8. 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.
9. Gan, C.K. & Donaldson, R.W. Adaptive Silence Deletion for Speech Storage and Voice Mail Applications. *IEEE Transactions on Acoustics, Speech, and Signal Processing 36, 6 (Jun. 1988)*, pp 924-927.
10. Heiman, G.W., Leo, R.J., Leighbody, G., & Bowler, K. Word Intelligibility Decrements and the Comprehension of Time-Compressed Speech. *Perception and Psychophysics 40*, 6 (1986): 407-411.
11. Hirschberg, J. & Grosz, B. Intonational Features of Local and Global Discourse. In Proceedings of the Speech and Natural Language Workshop, San Mateo, CA: Morgan Kaufmann Publishers, 1992, pp. 441-446.
12. Ju, S.X., Black, M.J., Minnerman, S. & Kimber D. Analysis of Gesture and Action in Technical Talks for Video Indexing. In *IEEE Trans. on Circuits and Systems for Video Technology.*
13. Kutik, E.J., Cooper, W.E. & Boyce, S. Declination of Fundamental Frequency in Speakers' Production of Parenthetical and Main Clauses. Journal of the Acoustic Society of America 73, 5 (1983), pp 1731-1738.
14. Lienhart, R., Pfeiffer, S., Fischer S. & Effelsberg, W. Video Abstracting, *ACM Communications*, December 1997.
15. Medan, Y., Yair, E. & Chazan, D. Super Resolution Pitch Determination of Speech Signals, *IEEE Transactions on Signal Processing, 39(1), Jan, 1991*, pp 40-48.
16. Merlino, A., Morey, D. & Maybury, M. Broadcast News Navigation Using Story Segmentation. In Proceedings of the 6th ACM international conference on Multimedia, 1997.
17. 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.
18. 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.
19. Resnick, P. & Varian, H.R. (Guest Editors) Recommender Systems. In *ACM Communications*, March 1997.
20. Silverman, K.E.A. The Structure and Processing of Fundamental Frequency Contours. Ph.D. dissertation, University of Cambridge, Apr. 1987.
21. Stanford Online: Masters in Electrical Engineering, 1998. http://scpd.stanford.edu/cee/telecom/onlinedegree.html
22. 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.
23. Stifelman, L. The Audio Notebook: Paper and Pen Interaction with Structured Speech *Ph.D. dissertation, MIT Media Laboratory*, 1997.
24. 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.
25. 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.
26. 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.