Audio Surrogation for Digital Video: A Design Framework

Kristin Boekelheide, E. Ashley Rogers Brown, Xin Fu, Gary Marchionini, Sanghee Oh, Gershom Rogers, Billy Saelim, Yaxiao Song, & Fred Stutzman

1. Introduction

Video content becomes increasingly important in WWW applications as the emerging global cyber infrastructure develops. Improvements in hardware (e.g., fast CPUs, graphics chips, inexpensive mass storage, inexpensive video cameras, cell phones), software (e.g., video editors, video player extensions to web browsers and other general-purpose applications, web development environments), and networking (steady increases in bandwidth, emerging quality of service schemes) enable much of the development in WWW-based video. These conditions support the current explosion in video content available online; meanwhile, television and movie producers vie to capture the emerging WWW-based video market. Most major content producers, networks and search engine companies have aggressive video retrieval and delivery services ranging from high-resolution file downloads to light-weight streams for mobile devices. At present, however, there are few innovative video retrieval systems that support sophisticated searching and browsing. Although there have been decades of content-based video retrieval research, today's commercial searching and browsing systems depend on text-based searching and simple query-by-example search interfaces. Some video digital libraries such as Open Video (www.open-video.org) and Fischlar (http://www.cdvp.dcu.ie/) offer more innovative search interfaces; however, much research and development remains to improve the effectiveness and efficiency of video search. Yang's (2005) study of how different user groups make relevance judgments for video content demonstrates the need for varied metadata and non-textual surrogates to find and filter video content.

Most of the video retrieval research and development efforts over the past decade focus on visual features (e.g., luminosity, color, texture, shapes) of video. This is understandable since the speech data in audio channels can be treated as text, enabling the application of traditional IR techniques. Five years of TREC Video results readily demonstrate the importance of linguistic data for retrieval; 2005 was the first year that some groups showed better performance with visual features than linguistic features (Over et al., 2005 http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5overview.pdf). However, given that other audio features are also important cues for meaning and retrieval, it may be useful to more fully include music and natural or artificial sounds in video retrieval systems. With better broadband access, using visual and audio data as entry points for retrieval is increasingly practical. Moreover, small form-factor devices with very limited screen real estate suggest the need for additional kinds of data cues. It is thus time to investigate audio surrogation as an important component of video retrieval and sense-making.

This paper provides a framework to guide audio surrogation research and development. It is meant to help system designers identify which kinds of audio surrogates are most appropriate for a specific system, and to help researchers develop research methodologies. After a brief review of the roles that surrogates play in retrieval and sense making, and of some characteristics of audio data, five types of audio surrogates are defined, potential applications are illustrated, and implementation issues are discussed. The paper concludes with a discussion of the implementation issues related to multiple kinds of surrogates in practical video retrieval systems.

2. Background for Audio Surrogation

2.1. Surrogates and Surrogates for Video

In general parlance, a "surrogate" refers to one thing that substitutes for another and is often used in the context of social relationships (e.g., a "surrogate mother"). In the information science literature, a surrogate is a condensed representation constructed to stand for a complete information object. The

retrieval literature demonstrates how people make sense from the surrogates or cues, which stand for the object so that the seeker can decide whether to retrieve it or not.

Whether designed to act as pointer to the original object (e.g., a result in a web search engine) or as a browsing tool (e.g., a film trailer produced to publicize an upcoming movie), all surrogates are created to acquaint users with the original object while reducing the time spent by a user. We refer to the ratio of real time run length of video to the amount of time a person spends to understand a surrogate as the compaction rate. Compaction rates for visual surrogates vary over two orders of magnitude, with the fast forward surrogates in Open Video offering a compaction rate of 64:1. User studies have demonstrated that much higher compaction rates are possible (Wildemuth et al, 2003). Bibliographic information and XML-encoded metadata are surrogates commonly used by people and machines. Text entries (e.g., keywords, titles, or abstracts) are traditionally used as surrogates because they are easily expressed in user queries and in system result sets (e.g., Borko & Bernier, 1975; Lancaster, 1991). Electronic systems permit new kinds of surrogation, and defining new kinds of surrogates for electronic objects is an active area of research in information science (e.g., Burke, 1999; Jorgensen, 2003; O'Connor, 1996) and signal processing (e.g., Jain & Hampapur, 1994; Kato, 1992; Lienhart et al., 1997).

During retrieval queries, surrogates enable decision-making by presenting search results in a uniform way, by supporting incidental learning and saving network capacity (Ding et al, 1999). Regardless of the kinds of surrogates a system provides to users, surrogates can be applied in distinct and creative ways in practice. Thus a good surrogate will have a primary use and many secondary uses. For example, all surrogates that are primarily for finding may also be pathways to understanding for experienced users.

The creation of surrogates is driven both by what kind of information a particular type of surrogate is intended to represent, and for what purpose (its function), as well as what a particular type of surrogate is capable of representing (its technical feasibility). The most basic dimension of intended usage ranges from 'find' to 'understand.' At one extreme, surrogates may be optimized for retrieval (e.g., a specific URL or ISBN), while at the other extreme they may convey substantial meaning (e.g., an abstract or critical review). Another important facet is level of abstraction. Surrogates may be concrete or visceral, depending mainly on perception and recognition. Surrogates may also be symbolic or noumenal, requiring significant amounts of cognitive processing. A third facet of surrogation is origination, which refers to whether the surrogate is extracted from the object (e.g., a keyframe) or originally constructed (e.g., keywords, titles, descriptions, or a customized film trailer). Origination is particularly important for automatic surrogation efforts. Finally, surrogates may be characterized by their physical formats, which include dimensions such as size, medium, and cost.

Although the literature on how people use surrogates and how indexers create them is quite rich for text resources, we are only beginning to understand what kinds of surrogates might be useful for videos. O'Connor was an early advocate of new kinds of surrogates for film, defining keyframes twenty years ago (O'Connor, 1985). Turner (1994) investigated the kinds of surrogates people found useful for finding video and Goodrum (1997) gathered empirical data on people using a small set of content-based and text-based surrogates to find video. The Informedia group (Christel et al., 1998; Christel & Warmack, 2001) created a series of surrogates (e.g., skims, collages) and conducted empirical studies to determine their effectiveness. Their user studies' results parallel the results of Ding et al. (1999) and recent results from the NIST TREC Video Track (TREC, 2003) that demonstrate the importance of textual and audio cues when combined with visual surrogates. The Open Video group has also created and evaluated a series of video surrogates. What is clear is that the electronic medium has greatly expanded the range and roles of surrogation (Marchionini, 1995).

Examples of non-textual surrogates that may be used for video retrieval include the following:

- Keyframe: a.k.a. “poster frame”; an image selected to represent the video, usually a single frame extracted from the video
- Storyboard: a.k.a. “filmstrip”; a set of images displayed in chronological order, usually in a tabular format
- Slideshow: a series of keyframes presented for viewing one at a time for a few seconds each, i.e. as if viewing a slideshow
- Collage: a dynamically-created, interactive image constructed of text and images from multiple videos, perhaps at different display sizes
- Video fast-forward: viewing a video with dramatically shortened frame rates, i.e., as if viewing a video on fast-forward
- Skim: a video clip created by compacting visual and audio information that summarizes a video, shortening viewing time while preserving the original frame rate
- Trailer: a pre-produced series of clips excerpted from a video
- Match bar: specific to video retrieval; uses a marked bar to indicate where query terms or images appear in a video
- Overview: a familiar figure or symbol that may include text and that represents one aspect of the video, e.g., timeline
- Preview: an introduction to a video; the term “preview” may be used as a synonym for any of the surrogate types listed above

With increasing use of digital video and rapid growth of accessible video collections, it is becoming increasingly important that potential video viewers are able to use abstractions to support sense-making and efficient browsing of video. Storyboards, fast forwards, and seven-second excerpts are examples of presentation schemes for abstractions of digital videos that have been used in the Open Video Repository. Each of these presentation schemes has its own strengths and limitations in representing videos. For example, storyboards and fast forwards are static and dynamic video abstractions respectively, but both neglect the audio components of video. While a seven-second excerpt preserves both video and audio information carried in a video segment, it only conveys selected information about the video: the seven seconds extracted from some part of the video can hardly represent the whole video.

The unique features of different surrogates can be selectively applied in various video retrieval systems. What is missing from the above set of surrogates are audio surrogates. In particular, audio surrogation can provide a tool to assist a user in understanding a video’s contents because it engages the user’s natural ability to hear and requires no training or additional efforts in sense-making. Thus, the goal here is to augment our video surrogation arsenal with audio surrogates.

2.2. Audio Perception and Sense Making

Because people are remarkably adept at recognizing exactly to what object or event they are listening, audio perception is a powerful stimulus for information gathering. Since the crafting of video is defined by the interplay between the visual and auditory channels, audio surrogates have the potential to be powerful tools for understanding the original contents of multimedia resources and for searching and browsing video resources. Many recent studies evaluate techniques for creating auditory, image, text or video abstractions of multimedia resources and test how well these abstractions carry the primary meaning of the original resources. Christel et al (1998) tested people’s information gathering and satisfaction with various kinds of audio, video and image skims. The spatial effect of auditory stimulus has also been studied. These studies evaluate how people browse the sounds around them and act differently depending on where the sounds come from (Arons, 1992; Stifelman, 1994; Kobayshi, 1997). However, despite study results indicating that people gain substantial understanding from the audio components of video, few practical examples of audio surrogates in current video retrieval systems exist.

Human auditory systems rely on a mixture of physical, perceptual, and cognitive processes. People hear by processing a series of vibration-causing waves that reach their ears. Since ears constantly monitor the surrounding audio environment, people are continually stimulated by complex, potentially confusing and sometimes redundant sounds- noises, voices, music, etc. However, humans differentiate sounds providing information from the mix of noises and process these as meaning. Hearing is the process of transforming sounds measured by physical parameters such as wave frequencies and pitch intensity to meaningful mental objects and events (Forrester, 2000).

According to the Model Human Processor (Card, et al,, 1983), when humans detect a stimulus, perceptual systems associated with working memory begin to process the stimuli. In the perceptual processing of auditory stimuli, people can recognize 5 letters within 100 msec on average, and it takes 1500 msec on average before the auditory memory information fades out of their working memories. This time actually ranges from 900 msec to 3500 msec, so the memory decay time varies and can be extended depending on various factors. For example, in the experiment of Schmandt & Mullins (1995), when people listened to a story in the 400 Hz channel and 100 msec tone setting, the attention decay time was 5000 msec; this was almost three times longer than the average time given in the Model Human Processor model. Interestingly, auditory memory decay time is even longer than that of visual stimuli. People can recognize 17 letters when they receive visual stimuli for 100 msec, three times faster than the letters recognized by the audio stimuli. However, the memory decay time of visual stimuli takes only 200 msec on average although it ranges from 70 to 1000 msec, seven times shorter than the audio stimuli. Thus, auditory recognition may be more effective for memorization and information understanding than visual stimuli.

Based on the information obtained from perceptual processing, cognitive systems interact with working memory and long-term memory in order to hold current information. This is useful in deciding how to respond to the stimuli, and in storing knowledge for future use. In the real world, the mind-stimuli process is much more complex. When hearing, people detect a mix of sounds simultaneously such as locations, harmonics, frequencies, continuity, and volume (Bergman, 1990; Wolfe et al., 2006). Based on information about the stimuli, people immediately distinguish the types of sounds as meaningful information or meaningless noise. In addition, people continue to see while listening; i.e., they sense both audio and visual stimuli continuously through multiple channels. Thus, the information captured based on the correlation between visual stimuli and audio stimuli is analyzed in the perceptual and cognitive systems, and these systems immediately, dynamically and repeatedly work toward sense-making and constructing knowledge.

Since there is a physical limitation of human listening capability to simultaneously capture and process multiple stimuli, the perceptual and cognitive loads in the listening process increase as the number of these channels increases (Stifelman, 1994). In addition, people listen selectively, focusing on one among multiple sound channels in order to reduce the cognitive effort and time required to listen and to increase the accuracy of their listening selection (Arons, 1992). People can easily recognize when sounds stop and shift their auditory attention toward other channels that are more interesting to them (Cohan, 1994; Schmandt & Mullins, 1995). Thus, for audio surrogates to become useful tools, it is necessary to find the proper levels of auditory stimulus and develop manipulated techniques that maintain people's attention long enough for them to understand the contents of the multimedia source.

Hearing is a mechanism through which people collect information from their ambient environment. Not only does hearing provide information about sounds themselves, it enables people to make decisions related to the sound's meaning, what kinds of objects or events are related to the sound, and where the sound is located.

3. Types of Audio Surrogates

3.1. Visual surrogates for audio

3.1.1. Overview of visual surrogates for audio. Visual surrogates for audio refer to various types of visual representations that can be created to display features of audio tracks. The overarching goal of designing visual surrogates for audio tracks is to help people gain a quick understanding of certain features of the tracks and make more informed audio manipulation or object retrieval decisions. There are several ways in which this goal may be achieved. First, visual surrogates can be used as a preview for audio tracks allowing users to make better judgments about whether the source object meets their need and should be downloaded. This would be useful on an audio/video search results page. Second, visual surrogates can act as a supplementary information channel when the audio tracks are being played. This is also predicted by multiple resource theory, and similarly by information redundancy (Gunther et al., 2004). The advantage is that people can look at the visual surrogates while listening to the audio clips. Moreover, if the user has a clear expectation on certain audio features, the surrogates can be used to provide interface controls that allow users to navigate to different parts of the video based on the audio features or to identify query match specification. To fulfill any of these goals, the surrogates must denote meaning at a glance. The representations must invite easy interpretation so that users learn how to read them without special training.

3.1.2. Applications and Approaches to visual surrogates for audio. Audio features that can be represented to help understand video contents include the following: types of sounds (e.g., environmental sound, human voice, animal voice, music, silence), types of speakers (e.g., gender, age, number of speakers), speaker alteration patterns, audio levels (loudness, excitement) and many other domain-specific features. A representative sample of visual surrogates discussed by our group is detailed below

(1) Bar chart

A bar chart can be used to denote the percent of time that each sound feature is present in an audio track thus allowing easy comparison between features. For example, in the left bar chart of Fig. 1, speech and music are two dominant features. In other words, during most of the time in this audio track, there is a person speaking while music plays. From the right bar chart, however, we can tell that there is quite a lot of white noise in addition to speech and music.

An advantage of using bar chart representation is that the sum of the audio features do not have to add up to one. Each type of sound is represented in relation to the entire audio track. For example, if a user seeks a video that features the sounds of WWII fighter jets, a bar chart will clearly indicate whether there is five minutes or 40 seconds of machine noise included in a given video. In a chart that shows machine noise as a percentage of all sounds on the video, the user would be uncertain as to the duration of the machine noise, seeing it only as a percent of the total length of all sounds. Additionally, more detailed information may be added to bar chart representation. For example, color coding could be used to denote male and female voices while sharing the same bar in the “speech” category.

A disadvantage of using a bar chart is that it only presents the total amount of time that a sound feature is present and not the chronological sequence of the sound features. For example, from either of the bar charts, we cannot determine whether the audio tracks start with music before going to speech, whether the two features alternate, or how frequently they alternate. In a video that teaches how to play guitar, the speech (verbal instruction and explanation) and the music (demonstration) may alternate much more frequently than in another video for a live guitar performance that features comments from the players even though the total durations were comparable.

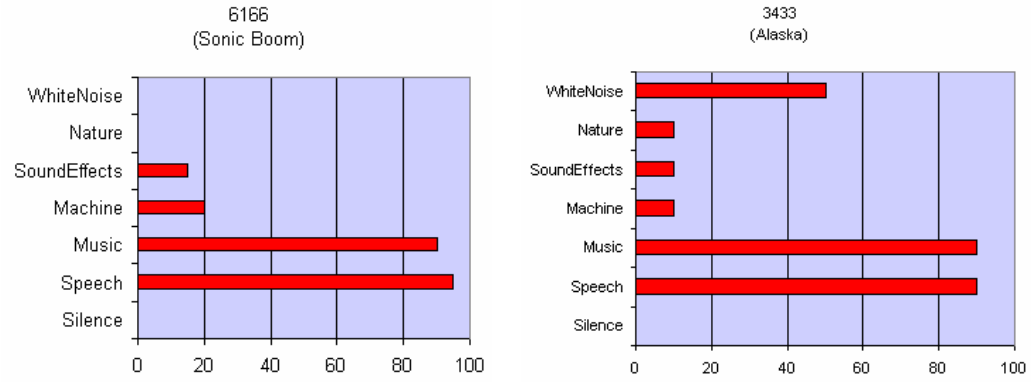

Fig. 1 Bar chart

(2) Pie chart

The pie chart is another way to show the breakdown of sound features within an audio track. Like a bar chart, a pie chart also allows easy comparison. We can also use cascading levels of the chart (such as pie of pie and bar of pie) to show finer grained categories (e.g., the pie chart on the right).

One difference between a pie chart and a bar chart is that the pie chart requires categories add up to one. In this sense, it is less flexible than a bar chart, however, the pie chart's simplicity may make them easier to quickly interpret 'at a glance.' One advantage of percentages adding up to one is that with further breaking down of the sound features, we can clearly tell the percentage of each feature and how the features overlap. However, with a bar chart, it is more confusing if we separate "speech with music" from "speech" and "music" categories). Moreover, like bar charts, a pie chart does not represent the temporal characteristics of the audio features, either.

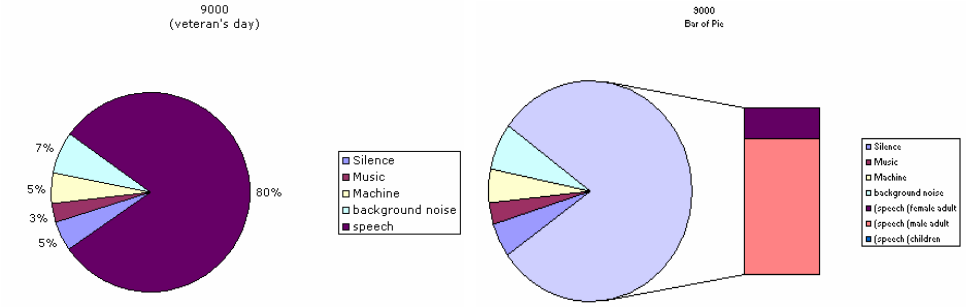

Fig. 2 Pie chart

Comment [FP1]: The two charts are not of the same size. Simply resizing it makes the text too small. Shall we draw the chart again using Excel?

(3) A linear sequence of colored blocks.

This approach is intended to preserve the temporal relationships between features while accurately representing a feature's overall length. A vertical bar with colored blocks represents the audio landscape of a multimedia object. Different colors represent different audio features, the alternation of colors represents the alternation of audio features, and the length of each single color block denotes the duration

of an audio feature. In the example in Fig. 3 (a), the bar corresponds to a 30-second audio track. After a brief silence, it starts with a short music sequence. The music is followed by a five-second speech and then another brief silence. After that, there is a long duration of speech, accompanied first by music and then briefly by machine sound. The audio fades out with a one-second silence at the end.

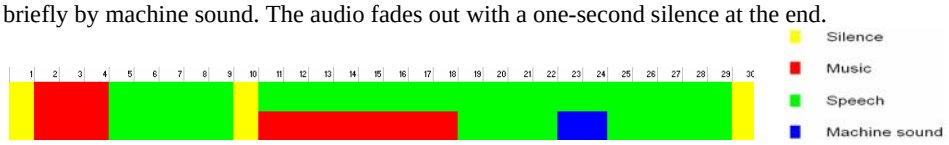

Fig. 3 (a) A linear sequence of colored blocks

Similar colors can be used to represent finer categories within sound types (e.g., speech may be broken down into female voices, male voices, and children's voices). Fig. 3 (b) represents a 60-second video, where all human speech is represented by green colors of different hues. The audio track represented here has frequent alterations of female and male voices, with a child heard only at the beginning and the end of the track.

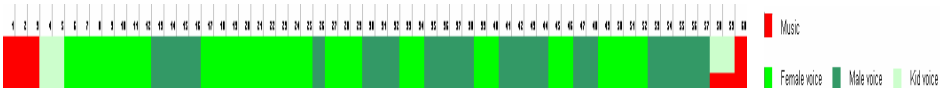

Fig. 3 (b) A linear sequences of colored blocks with finer grain

A challenge for this approach is to present the overlap or concurrency of features. In the above examples, the bars are "two features high" to represent feature concurrency, such as speech with background music. It is conceivable that the bars could become so dense with color and stacked sound features that they would be difficult to read quickly. Editing tools often display many concurrent channels but they are not meant for casual users. Another possibility is to use a blend of colors, but this solution also could become difficult to decipher.

We can also use the length of the bar to represent the length of a video so that the width of each color block is directly proportional to the duration of the corresponding feature. One challenge to these surrogates is depicting audio patterns for short and long videos as the visual scale must be adjusted for different video lengths, which may add additional load to users hoping to quickly get an overview of the video content.

3.1.3. Implementation requirements and challenges. A key step in creating the visual surrogate is the extraction of audio features. There have been studies on automatically classifying audio tracks into distinct audio classes. For example, Kimber and Wilcox (1996) applied hidden Markov models to classify audio signals into music, silence or speech. There are also domain-specific audio classification studies. For example, Li and Dorai (2004) used a support vector machine method to automatically classify instructional videos (including corporate education videos and professionally produced training videos) into seven audio classes: speech, silence, music, environmental sound, speech with music, speech with environmental sound, and environmental sound with music. The method was claimed to reach a 97.9% classification accuracy. A second challenge to visual surrogates for audio is that they take valuable screen real estate in the interface. A related challenge is that they compete for visual attention in a higly visual application like video. Nonetheless, with these existing audio classification techniques and potential new approaches, visual surrogates for audio can be implemented inexpensively and work as effective representations for videos.

3.2. Speech display of metadata (e.g., keywords, descriptions)

3.2.1. Overview of speech display of metadata. According to Gunther et al., 2004 "…auditory cues complement visual cues…by providing information redundancy…" (Gunther, 2004) Video excerpts, video fast-forwards, screen shots and storyboards are often supplemented by keywords, genre, and text descriptions.. Spoken metadata alongside visual surrogates allow the surrogates to be processed both auditorally and visually. Such redundancy may help identify video content. Additionally, "auditory information, when designed to complement the visual environment, is natural and people are innately comfortable with it - its use requires no training"(Gunther, et al., 2004). The video display formats in the Open Video Project, are available in silent form only with text based descriptions and keywords displayed beneath or next to the video screen shot. With the addition of spoken metadata a user would simultaneously view a video fast-forward, screen shot or storyboard with the addition of a spoken version of the keywords and/or description.

3.2.2. Applications and Approaches for speech display of metadata. Speech displays may appear in various formats: search results at mouse-over, audio keywords or descriptions during video fast-forwards, and storyboards. The audio keywords and descriptions are not necessarily separate options, but could be used in conjunction for audio surrogation (i.e., keywords spoken followed by description or vice versa.)

Spoken keywords or descriptions in search results upon mouse-over add audio to visual result list scanning. The result set includes a screen shot and the first two lines of a description with a list of all keywords. By mousing over the screenshot in the result list the user would be able to hear the entire description before clicking on the full record. The benefit of this option is that it provides aural information in an abbreviated version of the video record. The intuitive problem with the mouse over option is that result list scanning could be more time consuming. Also with this option the changing sounds when moving the mouse pointer through a result list might annoy the user.

Audio keywords or descriptions can also be used when viewing images or other silent video surrogates, such as fast-forwards, or when reading the spoken metadata, either human or text-to-speech with video surrogates. The spoken metadata could be initiated simultaneously with image loading or fast-forward playing. Alternately, sound controls could allow the user to have the options to turn the spoken metadata on or off for video fast-forwards.

Matching the duration times of audio tracks and video fast forwards is the primary challenge with the addition of spoken metadata. A lack of substantive keywords or descriptive text associated with some videos could potentially add little for the user. A corpus of similar videos with substantive but similar keywords or descriptive text would have a similar negligible effect. In instances where the metadata are both substantive and unique, the issue is one of timing. With a larger corpus of keywords affiliated with a short fast forward segment, spoken words may run beyond the length of the video fast-forward, while with less text, the user may hear spoken metadata for only a short portion of the fast-forward. Finally, adding spoken metadata to fast-forwards could be cognitively problematic with the lack of synchronization of ideas and images. If the complement of audio to video is redundancy, then the detriment is confusion caused by inserting new information unmatched by the fast-forward contents.

The addition of spoken metadata to storyboards would allow the user to review screen shots while hearing the keywords or descriptions. Time is less of a challenge when users are scanning the storyboard sections, unless the storyboards have few or many keyframes, short or long sets of keywords or other metadata may be repeated or trimmed to match typical storyboard scan times. By adding audio to storyboards, one might be able to capture some of the information offered in the motion of video fast forwards. The addition of audio descriptions to storyboards might additionally set a specific pace for scanning these shots.

3.2.3. Implementation requirements and challenges. Imperative to effective use of spoken metadata is the format of speech. Text-to-speech (text synthesis) and human speech are two options. The elements gained through synthesized voice are consistency, speed control and automatic generation; however, users may be more receptive to human voice. If human voice is chosen, recording time, speech cadence for consistency, and voice gender preference should be explored.

A challenge with all categories is the length and quality of current metadata. Not all videos are accompanied by descriptions. Existing keywords and descriptions should be examined for substance, length and match to video skims and storyboards.

Finally, cognitive load requires careful consideration. In the final results of the Gunther et al. examination of three-dimensional auditory cues, the results indicated that excessive use of aural information might cause a user to rely heavily on the audio and lower the attention given to the visual (Gunther, 2004).

3.3. Video/Audio Snippets and Skims

3.3.1. Overview of video/audio snippets and skims. A snippet is a short extract from a video or audio that is substantially shorter in time than the source video or audio. An skim is an aggregation of snippets and other surrogates that represent the entire video. In our case, an audio snippet is a short audio segment and an audio skim would consist of several audio snippets. With the rapid growth of accessible video collections, it is no longer practical for people to download and play the entire video/audio in order to determine its aboutness and relevance. Therefore snippets, being abstractions of the source video/audio, can be very useful in effective and efficient video/audio browsing.

3.3.2. Applications and Approaches to snippets and skims. A video skim is a temporal multimedia abstraction that incorporates both video and audio information from a longer source (Christel et al., 1998). Ideally, video skims can overcome the weaknesses of storyboards, fast forwards, and seven-second excerpts. With different compaction rates, a long video can be condensed to shorter skims of different lengths, e.g., a 30-minute video can be represented by a 3-minute skim with a compaction rate of 10:1. On the one hand, it is optimal to have compaction rates as high as possible; on the other, the higher compaction rates are, the less information we can comprehend from the skims. Therefore, we would like to achieve relatively high compaction rates while still being able to get a reasonable amount of information out of the videos.

Also, there is potential that people are able to get enough useful information from audio skims alone, especially in instructional videos with high concentrations of meaning in the verbal explanations. Audio skims are temporal multimedia abstractions that only incorporate the audio features of videos. The design of such audio skims requires that a balance be struck between compaction rate and usefulness of skims.

3.3.3. Implementation requirements and challenges for snippets and skims. A number of methods and techniques for creating video skims have been discussed in the literature [Christel, et al., 1998, Christel, et al., 1999, Yu et al., 2003]. Perhaps most notably, the Informedia Project created skims that were derived from seven different kinds of audio and visual features. In most cases, however, the creation and application of audio is rarely mentioned or minimally applied. One intuitive approach is to increase the frame rate for the entire video, which can also be extended to audio. As mentioned earlier, fast forwards without sound appear to be useful and content-rich video abstractions. The idea of “fast forwards with (or of) sound” is considered as a separate kind of audio surrogate below. We employ a simple approach that preserves the normal frame rate while still greatly reducing viewing time. Here we consider the key design parameters for audio snippets that play at normal frame rates.

The most direct approach is to sample small audio segments at intervals across the entire video. If we aim to achieve a certain compaction rate, the design decisions revolve around the length of the audio snippets (e.g., 3 seconds, 4 seconds, or n seconds); the sampling rate for the skims (e.g., every 30 seconds, 60 seconds, or 120 seconds); different subsampling techniques (e.g., starting from the first "good" audio skim, or starting at the beginning of the video); and rules for selecting "important" video and audio components. Though we can play the audio skims at 1.5 to 2 times speed to further increase the compaction rate, it is beyond the scope of this paper.

We conducted a pilot study where we create video or audio skims by subsampling the original videos at certain lengths of video/audio snippets and at fixed intervals (i.e., taking the first n seconds of each N second intervals of the video). For example, with a snippet length of 3 seconds and a sampling rate of 30 seconds, the resulting video/audio skims comprise seconds 0-3 of the source video, followed by seconds 30-33, seconds 60-63, and so on. The sampled snippets are then concatenated (to form skims) and played back at the original frame rate.

We denote the snippet length as *n* and the sampling rate as *N*. Apparently, the greater *n* is, the more consistent and useful the skims are. Similarly, the more samples we keep from the original video (that is, the smaller *N* is), the better the snippets we get. However, given a fixed compaction rate, the snippet length *n* is proportional to the sampling rate *N*, thus inversely proportional to the number of samples (snippets). With very large *n* and *N*, each snippet sample is consistent, but the information conveyed by the entire skim can be partial because of the small number of snippet samples. The extreme case of this is an n-second excerpt, where a single piece of the video represents the entire video. By way of contrast, with very small *n* and *N*, each track sample is so "choppy" that the concatenated skim is useless, even though a large number of snippet samples have been extracted. An extreme case of such choppiness can be observed during rapid playback of the video (with sound) or audio.

In our pilot experiment, we used *n = 2, 3…5* and *N = 30, 60, or 120* depending on durations of the original videos/audios. The experimental results show that people are able to determine the "aboutness" of the videos from their video/audio skims. Generally, *n = 2* sec was found to be too short and resulted in very "choppy" snippets, while *n=3* sec was acceptable, and *n=5* sec produced very good snippets.

It was also observed that given the same amount of "preview" time, audio skims alone (with no visual cues) convey much less information than video skims. One advantage of audio skims over their video counterparts is that audio skims are more cost-effective in terms of storage space and bandwidth requirements. However, in an era when mass storage has become increasingly less expensive, this advantage seems to be less of a consideration.

One scenario where audio snippets and skims may be useful is when video skims or other visual surrogates are not supported by the systems used for video searching and browsing, such as small devices like low-end cell phones and early models of digital music players. For example, users may be interested in searching and browsing videos using small devices while commuting between work and home. Furthermore, even for cell phones or newer models of digital music players that can play videos, downloading and storing large files can be quite problematic, given the limited bandwidth and storage space supported by such devices. Therefore, audio snippets and skims, being smaller in size and informative in content, can become very useful in a portable device context.

3.4. Compressed audio

3.4.1. Overview of Compressed audio. Time compression, in the context of audio or video surrogates, refers to a process through which the surrogates are shortened without the loss of content. Simplistically,

it is the “speeding-up” of a surrogate. According to Omoigui et al., (1999) there are two prominent types of time-compression, linear time compression and skimming. In this section, we concentrate on linear time compression (speeding up) rather than skimming. The challenge of time compression is twofold. First, a digital process must be undertaken to compress the surrogate content into a shorter time sequence. Second, the content must be normalized for pitch so that the compressed audio is not pitched abnormally. In doing so, the end result of the time compression process is a faster version of the audio surrogate that is human-intelligible.

A review of time compression literature shows publications dating back to the 1960’s. The fundamental issues of time compression revolve around a few key questions. First, the question of pitch-control has led to the development of a number of algorithmic techniques, as explored by Arons. (Arons, 1992). The second question involves the optimal rate at which humans can intelligibly perceive time-compressed communications. As evidenced by Schwab & DeGroot (1993), Omoigui et al. (1999), and Vemuri, et al., (2004), the optimal rate at which humans can perceive time-compressed audio is highly variable, though conditionally established. Vemuri’s work reflects the maturity of this particular line of inquiry, exploring the rates at which different populations can successfully interact with time compression. Finally, use cases and effects of time compression are explored by Vemuri et al., (2004) and Arons (1992).

3.4.2. Applications and Approaches to compressed audio Arons (1992) conducted a thorough review of the lineage of time compression. In the 1950’s through 1970’s, the domain of time compression was primarily limited to alteration of playback speeds, generally in the research context of assisting the visually handicapped. Early time compression techniques included human subjects speaking rapidly, speeding up analog tape playback, speech synthesis, vocoding and silence removal (Arons, 1992). Eventually, the domain of time compression moved towards an algorithmic, sampling-based approach. At a very basic level, sampling involves the removal of audio sequences that do not contribute substantially to the overall message. Techniques of sampling vary, but a simple example might be the shortening of long consonants. If a consonant is drawled, the sampler can effectively remove the less-necessary part of the consonant, while maintaining the intelligibility of the spoken word.

Arons goes on to describe a number of sampling techniques. These techniques were primitive in the 1970’s, but they continue to act as the basis for our sampling practice today. Brute-force methods of sampling include interrupted signal, Fairbanks sampling and dichotic presentation (Arons, 1992). In interrupted signal, the most primitive sampling technique, splices of sound are removed at a constant interval. Through application of Fairbanks sampling, those splices are reassembled to create a shortened audio product. With dichotic presentation, an audio source is interrupted (Interrupted signal sampling) and two channels are extracted. These two channels are then started at a time offset to a listener’s left and right auditory channel. The offset ensures time compression without loss of signal (Arons, 1992). Eventually, these sampling technologies evolved into selective sampling, the common form of sampling in use today. With selective sampling, an algorithm decides which segments of the audio track it can most effectively remove (Arons, 1992).

3.4.3. Implementation requirements and challenges A problem common to sampling techniques is the inability to account for pitch modulation. When language is sped up, an unwanted consequence is the unnatural raising of pitch (Arons, 1992; Omoigui, et al., 1999; Vemuri, et al., 2004; Schwab & DeGroot, 1993). In creation of the Synchronized Overlap Add Method (SOLA), Roucos and Wilgus developed an algorithmic method for time compression that did not noticeably affect signal pitch (Arons, 1992). Emerging as an industry standard, the SOLA method of time compression has gone through a number of iterations. As recently as 2004, Vemuri et al. (2004) used the SOLAFS method (a newer type of SOLA compression) in their studies.

Time compression is used commonly; many times, we may not even notice our interaction with time-compressed materials. One example we have all experienced is time compression in television commercials (Schwab & DeGroot, 1993), particularly in car commercials where the "fine print" is spoken. Time compression exists in a number of different domains. Schwab & DeGroot ((1993) point to its uses in instructional services and interactive voice response systems. Vemuri et al., 2004 documented the use of time compression in transcription services, particularly in voice-mail transcription services. Arons (1992) explores the use of time compression in materials designed to help the visually impaired. With the creation of SpeechSkimmer, Arons creates a system for the facilitated search of audio documents (Arons, 1997). In pursuit of the Memex, a team from Microsoft designed a time compression system with the explicit goal of reviewing computer-generated transcripts (Omoigui, et al., 1999).

According to Vemuri et al., (2004) the average speech rate of a human being is 180wpm, while the average reading comprehension rate of a human is 400wpm . Obviously, there is a gap between our ability to communicate via spoken words and our ability to comprehend them. In engaging the user with time-compressed audio, the central question involves the balance between compression and intelligibility. In Schwab & DeGroot (1993) compaction rates between 10 percent and 30 percent were explored, with older subjects displaying less ability to comprehend time-compressed audio. Vemuri et al., (2004) explored the ability of individuals to comprehend time-compressed text with the assistance of transcripts. At rates of 350wpm spoken, the subjects showed up to a 90 percent comprehension rate with the assistance of transcripts. At rates of 450 wpm spoken, the subjects displayed rates of up 62 percent comprehension. These compaction rates for compressed speech offer quite modest compaction rates in the 2:1 or less range.

In studies, the core challenge that confronts the researcher is determining the correct rate at which to present time-compressed material. In doing so, the researcher must determine the optimal time compression rate, as it corresponds to intelligibility. Age and experience with the time-compressed language strongly affects the optimal rate. In the development of SpeechSkimmer, Arons implemented a device that allowed flexible control over rates of time compression (Arons, 1997). This goal continues to inform the central research focus of time compression. Algorithms such as SOLAFS can time-compress audios without introducing noticeable pitch disturbance, and thus determining subjects' rightness-of-fit in time-compression speed is a research goal. Obviously, rightness-of-fit is highly circumstantial; subjects experiencing simple, native language ("The dog is under the car") may be able to process a larger portion of information at high speeds, compared with more complicated language. In the context of video search, where time compression could assist searchers, an optimal time-compression rate could be hypothesized after analysis of a number of factors. These factors include:

- Subjects' experience with the corpus
- Amount of material presented to the subject
- Acceptable comprehension rate
- Subject language experience (e.g., native speaker)
- Importance of accuracy in comprehension (along with understanding the words, does the subject also comprehend the message?)

Analysis of these factors could lead to an acceptable time-compression rate for video surrogates. If the subjects are to be presented with unfamilar material, where they have to make high-accuracy judgments, the optimal compression rate could be as low as 10 percent. If the subjects are to be presented with familiar material, where accuracy in comprehension is less important, the optimal compression rate could be greater than 30 percent. In general, user studies like those managed by Venturi and Omoigui lay a framework for discovery of the optimal time compression rate.

3.5. Parallel Audio Streams

3.5.1. Overview of parallel audio streams. The very act of our listening, if considered an interface, embodies a highly sophisticated function: the capability to extract information from several audio channels simultaneously. This capability is most evident in the presence of multiple concurrent conversations (even while holding one of our own), in which we attend selectively to information from multiple sources, as we might do while attending a cocktail party. For an interface generally regarded as "slow and serial" (Stifelman, 1994), the phenomenon of simultaneous listening as an opportunity to compress listening time merits serious consideration. In fact, many applications would benefit from a way to harness such a capability, from auditory browsing of wearable devices (Schmandt and Mullins, 1995), to understanding of a multimedia collection (e.g., The Open Video Project).

3.5.2. Applications and approaches to parallel audio streams. To induce a "cognitive" phenomenon through human computer interaction, one must at the very least understand how the phenomenon is manifested. An extensive account of previous findings, all vigorously supported by experiments, is given by Stifelman (1994). Among the most notable findings are:

3.5.2.1 In the presence of two channels, one tends to be more attentive to the primary channel (the one which is "consciously" selected), whether or not "shadowing"—where the listener also repeats the speech verbatim—is used (Treisman & Geffen, 1967). Furthermore, listening to target words in the secondary channel is regarded as disruptive to listening to the primary channel.

3.5.2.2 In the presence of two channels, unattended materials are retained in short-term memory for no more than five seconds (Norman, 1976). Such is the case regardless of shadowing, and explains why it is possible for us to be engaged by the information in the unattended channel.

3.5.2.3 In the presence of more than two channels, Stifelman reported statistically significant degradation in the performance of listening comprehension, as the number of channels went from one to three.

3.5.2.4 A number of strategies have been proposed to address the third finding, such as the spatial segregation of channels, the segregation of channels by frequency ranges, the segregation of channels by pitches, and the staggering of the onset of different messages. Experiments with the strategies in various combinations, however, show mixed results. For example, it takes time for users to locate spatially distinguished channels – a major distraction if the messages carried by those channels are judged to be irrelevant later. On the other hand, Spieth et al. (1954) showed that by deviating the onset time of simultaneous messages, errors in listening are reduced.

Schmandt and Mullins (1995) used several of the findings above to guide the design of their prototype, AudioStreamer. In particular, a head related transfer function (HRTF) was used to segregate the audio channels spatially. To shorten the time needed for locating channels spatially, head movements were used to indicate channel preferences. Furthermore, channels were separated by an angular distance that was neither too small for them to be perceived distinctively, nor too big for them to be located quickly. "Salient events" in messages were identified and marked in advance, based on acoustic cues and closed caption information, so that users could be alerted to them during the playback of the messages.

3.5.3. Implementation requirements and challenges. Clearly, without any prior knowledge of users' preferences, as is often the case with auditory browsing, it would be impractical to define salient events in advance. In an attempt to gain some firsthand experience with auditory browsing, we informally set up three laptops, each playing a different audio downloaded from The Open Video Project, at different locations in a room, and solicited comments from a group of about ten graduate students. Almost

unanimously the group considered the interference of channels to be so severe that the audio was utterly undecipherable.

Why does simultaneous listening work in some cases, but not in others? As of now there is not a clear answer. However, some insights may be gained from a similar issue pondered by Stifelman (1994): "whether users serendipitously overhear information of interest in a secondary channel while focusing on the primary channel, or they simply monitor each channel serially." If the latter is the case, the interference experienced in our experiment could be disruptive to such serialization. In retrospect, a theory of divided attention on which a study of simultaneous video was based (Slaughter et al., 1997), known as multiple resource theory, may be pertinent here. Not only does the theory state that the resources available to mental tasks are in dichotomous dimensions, but it also predicts better performance if dual tasks are split between two different sources (auditory, visual), than if they are split within one source (two visual inputs). The findings of Slaughter et al. (1997) provide empirical support both to the theory and to the premise of limited resources available for divided attention, by showing a significant decline in performance when the number of video surrogates shown simultaneously to users increased from two to four. Although the findings do not automatically lend themselves to an explanation of why simultaneous listening fails, their parallels with 3.5.2.3 suggest that the same mechanism underlies both visual and auditory comprehension.

Notice also that the findings above focus on the receiver role of a simultaneous listener. Could it be the case that, in a "cocktail party" like scenario where the audio tends to be conversational, that simultaneous listening is merely the result of some greater collective awareness of competing audio? It is certainly not our intention to rule out serendipity. Quite the contrary, we believe serendipity is the very foundation of auditory browsing. It may well be the case that the "collective awareness" mentioned above manifests serendipity, and that it is through collective awareness that the stabilization of concurrent conversations is attained. Nonetheless, we know too little about serendipity for audio, other than the fact that it is upper-bounded by a window of short-term memory that cannot exceed five seconds.

4. User Interface Design Challenges

As with any design framework, it is important to consider user profiles and user tasks as design decisions are made, as well as the setting in which the use of an information retrieval interface might be evaluated. However, as Robertson (2000) observes, it is often problematic to evaluate the effectiveness of retrieval interfaces, because it is difficult to: 1) characterize the retrieval tasks that a user might wish to perform; 2) determine whether a particular task was completed "successfully," and; 3) choose a single interface design, given multiple interface design possibilities. According to Smeaton (2000), three key ideas have driven the design of non-text (multimedia) content-based retrieval systems: 1) such systems work best when domain-specific; 2) automatic extraction tools have been very difficult to develop, resulting in considerable reliance on interactive tools that require user input, and; 3) such systems work best for simple tasks that can be performed consistently by users, as opposed to more complex tasks where outcomes are more difficult to measure. Furthermore, given the relative paucity of studies that focus on audio surrogation, it is a significant challenge to develop an interface that employs audio surrogates as a means of assisting users with the retrieval of multimedia content.

There are design challenges associated with each of the five techniques described in this paper (visual surrogates for audio, speech display of metadata, video/audio snippets and skims, compressed audio, and parallel audio streams). For instance, the automatic extraction of audio features from videos tends to work best with domain-specific videos. Even when the audio extraction is highly accurate, it is far from certain that a user would likely use an audio surrogate such as a bar chart or a linear sequence of colored blocks than another type of surrogate, such as a video thumbnail or keyframe. One of the techniques some

scholars view as a remedy for many of the video and audio surrogation shortcomings is the use of controls such as sliders, although designers must be careful not to add additional time to usage, which defeats a major purpose for the surrogates to begin with. For instance, Hürst, et al., (2004) describe an "elastic audio slider" that is particularly well-suited for the playback of compressed audio. The elastic audio slider is an interface component that facilitates slow and fine scrolling, as well as rapid navigation, depending on the distance between the slider thumb and the mouse pointer. Richter et al. (1999) also advocate for the use of sliders, and they propose the use of multiple timelines, each of which has its own slider, where the focus area on the topmost timeline determines the range that is displayed on the timelines beneath it, an approach that they tested on the playback of long media streams.

In addition to the five techniques described in this paper, researchers have developed numerous ways to characterize, browse and otherwise interact with audio. Schmandt (1998) describes what he calls "braided audio," in which each distinct sound is sequentially amplified so that it momentarily sounds more dominant to the listener. Schmandt sees braided audio as analogous to a visual collage that combines segments from multiple images. He describes the use of braided audio within the context of an Audio Hallway that is intended for browsing collections of related audio files. However, as Schmandt points out, presentation of audio files for browsing is complicated by the fact that audio is both serial and transitory—that is, for video, it is possible to render a thumbnail or a similar snapshot view of the video, while for audio, rendering something comparable to a thumbnail (that is, visual surrogates for audio) is still in the investigative stages. As a means of solving this problem, Ranjan (2005) describes an interface that displays both video and audio surrogates, where PowerPoint slides are shown as thumbnails in conjunction with an audio transcript timeline, and where "multi-focus zooming" is employed to enable the user to focus on more than one section of the audio transcript at a time.

Determining what audio surrogates to make available is closely related to determining the most suitable type of browser interface for audio surrogate display. Lee et al. (2000) describe six browsing interfaces for video, including a Timeline Bar Browser that makes it possible for the user to move a control and thus to change the set of keyframes that is displayed, and a Dynamic Overview Browser that causes detailed keyframes to display upon mouseover on a given keyframe. Hürst and Stiegeler (2002) observe that many user interfaces which are intended for multimedia tend to function much as a VCR does, with buttons that are analogous to standard VCR functions, along with some type of scrollbar-like control that is intended to help the user skip forward or back. Hürst and Stiegeler describe an interface that adds controls to enable users to zoom in or out when scrolling and thus to change the scrolling speed, a feature that is particularly useful when fine control is required in order to find a particular section of a multimedia file.

Although individual use of the five types of audio surrogates discussed in this paper might not produce measurable improvements in user retrieval success, various combinations of the five types might produce more noticeable improvements. Ranjan (2005) advocates for what he calls a hybrid technique that can leverage the advantages of multiple approaches and thus achieve better results. Vemuri et al. (2004) demonstrated the potential of a hybrid approach by combining time-compressed audio and transcripts in a single interface. The interface that Vemuri et al. created enabled users to both display audio transcripts and to adjust the speed of the audio playback.

An area that might benefit from further research as it relates to audio and video surrogation is the development of "context-aware applications." According to Dey (2001), a context-aware application can provide three types of operations for users: 1) present information and services; 2) automatically execute services, and; 3) correlate context with information, to support retrieval. Dey goes on to describe a Context Toolkit that is intended to store information about context (historical information about user actions) and to aggregate that contextual information for use by other applications.

5. Conclusion

Linguistic data in video is extremely useful for retrieval and sense-making, and these data have mainly been implemented in text representations while most research and development has focused on visual feature representations and analysis. Additionally, other kinds of audio information can help people quickly understand whether it is worth looking further or downloading a full video. This paper provides the beginnings for a framework for audio surrogation. There are three interdependent primary factors that characterize these surrogates: compaction rate, user effort, and system requirements.

Compaction ratio is the simplest factor in that it represents a kind of return on investment for people searching for video. A very high compaction ratio will save a lot of time in making relevance judgments, but those judgments may not be accurate or complete. Three of the audio surrogates discussed here offer potentially high compaction rates (10:1 or more): visual displays for audio features, spoken metadata, and snippets or skims. Compressed audio and parallel audio streams offer relatively low compaction rates.

User effort is a complex factor that has several subfactors. Three subfactors of importance are the number of perceptual system channels stimulated by the surrogate, the perceptual load required in these channels, and the level of abstraction of the surrogate. Divided attention theory and similar psychological theories suggest that involving multiple channels can be useful if these channels are coordinated. The amount of perceptual attention required in each channel will surely influence surrogate usefulness, especially if multiple channels are involved because the incoming signals must be interpreted and integrated by cognitive processes. Level of abstraction refers to how much decoding must be done to turn perceptual signals into meaningful concepts. For the surrogates considered here, visual surrogates for audio are the only ones that involve multiple perceptual channels. These surrogates require high levels of visual load, especially since other visual information will also be active in the user interface. They also require decoding and thus overall require high levels of user effort. Spoken metadata require moderate amounts of human audio channel bandwidth and moderate amounts of decoding, depending on the kinds of metadata spoken (e.g., keywords will require more decoding than sentence descriptions in natural language). Audio skims require moderate levels of perceptual bandwidth but may require high levels of decoding as several snippets are brought together from different parts of the video. Compressed audio requires high perceptual bandwidth to catch words, melodies, or other sounds, but generally requires low levels of decoding, unless the sounds are themselves symbolic. Finally, parallel audio requires very high perceptual bandwidth but may not require much decoding.

System requirement have two kinds of subfactors: the cost of creating the surrogates and the costs of (dis)playing the surrogates. Visual surrogates for audio are relatively easy to create but require much screen real estate to display. Spoken metadata is easy to create and play. Skims are of medium cost to create and medium difficulty to play and control. Compressed audio is of medium cost to create and requires modest play and control costs. Parallel audio is relatively easy to create from a system point of view but requires high user interface control costs on the part of users.

This work aimed to identify a small set of audio surrogates and characterize their potential for use in video retrieval systems. In some settings, audio surrogates will stand alone, in most cases they will be combined with other kinds of surrogates. The different audio surrogates offer different advantages and disadvantages with respect to amount of user effort, costs of creation and display, and overall compaction rate. This paper considers some of the design issues involved in audio surrogation and offers developers a perspective on the many design tradeoffs necessary in incorporating audio surrogates into video retrieval systems.

6. References

Arons, B. (1997) SpeechSkimmer: A System for Interactively Skimming Recorded Human Speech. ACM Transactions on Computer Human Interaction, 4(1), 3-38.

]Arons, B. (1992) Techniques, Perception and Applications of Time-Compressed Speech. Proceedings of American Voice I/O Society 1992 Conference, 169-177.

Bergman, A. S. (1990). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press.

Borko, H. & Bernier, C. 1975. *Abstracting Concepts and Methods*. NY: Academic Press.

Burke, M. (1999). *Organization of Multimedia Resources*. Hampshire, UK: Gower Publishing.

Card, S. K, Moran, T. P, & Newell, A. (1983). The Psychaology of Human-Computer Interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Christel, M. Smith, C.R. Taylor, and D. Winkler, "Evolving Video Skims into Useful Multimedia Abstractions", Proc. CHI '98, ACM, New York, 1998, pp. 171-178.

Christel, A. Hauptmann, A. Warmack, and S. Crosby, "Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library", Proc. IEEE Advances in Digital Libraries Conference, Baltimore MD, 1999.

Christel, M et al. "Collages as Dynamic Summaries for News Video", Multimedia '02, ACM: 2002. pp. 561-69.

Christel, M et al. "Multimedia abstractions for a digital video library", ICDL, Proceedings of the Second ACM International Conference on Digital Libraries: 1997. pp. 21-29.

Cohen, J. (1994). Monitoring background activities. Auditory Display: Sonification, Audification and Auditory Interfaces. Reading, M.A: Addison-Wesley.

Dey, A. (2001). Understanding and Using Context. Personal and Ubiquitous Computing, 5(1), 4-7. Retrieved April 5, 2006 from http://diuf.unifr.ch/pai/education/2002_2003/seminar/winter/ubicomp/PeTe5-1.pdf

Ding Wei, Marchionini Gary & Dagobert Soergel. "Multimodal Surrogates for Video Browsing", International Conference on Digital Libraries, Proceedings of the Fourth ACM Conference on Digital Libraries: 1999. pp. 85-93.

Forrester, M. A. (2000) Auditory perception and sound as event: theorizing sound imagery in psychology. Sound Journal. Retrieved on March 31, 2006 from http://www.kent.ac.uk/sdfva/sound-journal/forrester001.html

Goodrum, A. (1997). *Evaluation of Text-Based and Image-Based Representations for Moving Image Document*s. Unpublished doctoral dissertation, University of North Texas.

Gunther, R., Kazman, R, and MaccGregor, C. (2004) Using 3D sound as a navigational aid in virtual environments. Behaviour and Information Technology. 23(6), 435-446.

Hürst, W., Lauer, T., and Götz, G. (2004). An Elastic Audio Slider for Interactive Speech Skimming. Proceedings of the third Nordic conference on Human Computer Interaction (NORDCHI 2004), 277-280.

Hürst, W. and Stiegeler, P. (2002). User Interfaces for Browsing and Navigation of Continuous Multimedia Data. Proceedings of the second Nordic conference on Human Computer Interaction (NORDCHI 2002), 267-270.

Jain, A. & Vailaya, A. (1996). Image retrieval using color and shape. *Pattern Recognition*, 29(8), 1233-1244.

Jorgensen, C. (2003). *Image Retrieval: Theory and practice*. Scarecrow Press.

Kato, T. (1992). Database architecture for content-based image retrieval. *Image Storage and Retrieval Systems: Proceedings of SPIE* (vol 1662). 112-123.

Kimber, D., & Wilcox L. (1996). Acoustic segmentation for audio browsers. Proceedings of Interface Conference, Sydney, Australia, July 1996.

Lancaster, F. (1991) *Indexing and Abstracting in Theory and Practice* (3rd Ed). Champaign, IL: University of Illinois Press.

Lee, H., Smeaton, A., Berrut, C., Murphy, N., Marlow, S, and O'Connor, N. (2000). Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video. Proceedings of the Fourth European Conference on Digital Libraries, Lisbon, Portugal, September 2000 (ECDL 2000).

Li, Y. & Dorai C. (2004). SVM-Based audio classification for instructional video analysis. ICASSP 2004.

Lienhart, R., Pfeiffer, S., & Effelsburg, W. (1997). Video abstracting. *Communications of the ACM, 40*(12), 54-63.

Marchionini, G. (1995). *Information Seeking in Electronic Environments.* Cambridge, England: Cambridge University Press.

Marchionini, G., & Geisler, G. (2002). The Open Video digital library. *D-Lib Magazine*, *8*(12). http://www.dlib.org/dlib/december02/marchionini/12marchionini.html.

Norman, D. A. (1976). *Memory and attention :An introduction to human information processing* (2d ed.). New York: Wiley.

O'Conner, B. (1985). Access to moving image documents: Background concepts and proposals for surrogates for film and video works. *Journal of Documentation, 41*(4), 209-220.

Omoigui, N., He, L., Gupta, A., Grudin, J. and Sanocki, E. (1999) Time-Compression: System Concerns, Usage, and Benefits. Proceedings of CHI 1999, 136-143.

Ranjan, A. (2005). Browsing Archived Meeting Audio and Time-Synchronized Data. Master of Science Thesis, Department of Computer Science, University of Toronto. Retrieved April 13, 2006 from http://www.dgp.toronto.edu/~aranjan/Docs/Abhishek_MSThesis_2005.pdf.

Richter, H., Brotherton, J., Abowd, G., & Truong, K. (1999). A Multi-Scale Timeline Slider for Stream Visualization and Control. GVU Center, Georgia Institute of Technology, Technical Report GIT-GVU-99-30, June 1999. Retrieved April 13, 2006 from http://www-static.cc.gatech.edu/fce/eclass/pubs/tech/GIT-GVU-99-30.pdf.

Robertson, S. (2000). Evaluation in Information Retrieval. European Summer School in Information Retrieval (ESSIR 2000), Varenna, Lago di Como, Italy, 81-92.

Slaughter, L. A., Shneiderman, B., & Marchionini, G. (1997). Comprehension and object recognition capabilities for presentations of simultaneous video key frame surrogates. *ECDL*, 41-54.

Schmandt, C. (1998). Audio Hallway: A Virtual Acoustic Environment for Browsing. Proceedings of ACM Symposium on User Interface Software and Technology (UIST 98), 163-170.

Schmandt, C., & Mullins, A. (1995). AudioStreamer: Exploiting simultaneity for listening. CHI '95: Conference companion on human factors in computing systems, Denver, Colorado, United States, 218-219. from http://doi.acm.org.libproxy.lib.unc.edu/10.1145/223355.223533

Schwab, E. and DeGroot, J. (1993) Listener Response to Time Compressed Speech. Proceedings of CHI 1993 Conference Companion on Human Factors in Computing Systems, 81-83.

Smeaton, A. (2000). Indexing, Browsing and Searching of Digital Video and Digital Audio Information. Tutorial Notes, European Summer School in Information Retrieval (ESSIR 2000), Varenna, Lago di Como, Italy, 93-110.

Spieth, W., Curtis, J. F., & Webster, J. C. (1954). Responding to one of two simultaneous messages. *The Journal of the Acoustical Society of America, 26*(3), 391-396.

Stifelman, L. (1994). The cocktail party effect in auditory interfaces: A study of simultaneous presentation, MIT Media Lab.

Treisman, A. M., & Geffen, G. (1967). Selective attention: Perception or response? Quarterly Journal of Experimental Psychology, 19, 1–18

Turner, J. (1994). Determining the subject content of still and moving image documents for storage and retrieval: An experimental investigation. Unpublished doctoral dissertation, University of Toronto.

Vemuri, S., DeCamp, P., Bender, W. and Schmandt, C. (2004) Improving Speech Playback Using Time-Compression and Speech Recognition. Proceedings of CHI 2004, 295-302.

Wildemuth, B. M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., & Gruss, R. (2003a). How fast is too fast? Evaluating fast forward surrogates for digital video. Paper presented at the ACM/IEEE Joint Conference on Digital Libraries, Houston, May 2003.

Wolfe, J. M., Kleunder, K. R., & Levi, D. R. (2006). Sensation & Perception. Sinauer: Sunderland, M.A. The practice website is retrieved on March 31, 2006

Yang, M. (2005). An exploration of users' video relevance criteria. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.

B. Yu, W. Ma, K. Nahrstedt, and H. Zhang, "Video Summarization based on User Log Enhanced Link Analysis", MM'03, Berkeley, CA, 2003.