Dynamic Soundscape: mapping time to space for audio browsing

Minoru Kobayashi[†*1] **and Chris Schmandt**[†]
minoru@media.mit.edu, geek@media.mit.edu

[†] Speech Interface Group, MIT Media Laboratory
E15-252, 20 Ames St., Cambridge, MA 02139

[*] NTT Human Interface Laboratories, 1-2356 Take,
Yokosuka, Kanagawa 238-03 Japan

ABSTRACT

Browsing audio data is not as easy as browsing printed documents because of the temporal nature of sound. This paper presents a browsing environment that provides a spatial interface for temporal navigation of audio data, taking advantage of human abilities of simultaneous listening and memory of spatial location. Instead of fast-forwarding or rewinding, users browse the audio data by switching their attention between moving sound sources that play multiple portions of a single audio recording. The motion of the sound sources maps temporal position within the audio recording onto spatial location, so that listeners can use their memory of spatial location to find a specific topic. This paper describes the iterative design approach toward the audio browsing system, including the development of user interface devices.

Keywords

audio browsing, spatialized audio, simultaneous listening, selective listening, spatial memory

INTRODUCTION

When browsing printed documents, we rapidly shift our focus of attention to quickly skim their contents. We recognize the size and structure of the document, and use our visual spatial memory to recall and search for specific topics. Browsing audio is not so easy. When browsing an audio recording, we must repeatedly play and skip portions; without playing, we cannot perceive the sound or its contents. We must hear all of the audio stream to reliably capture all its topics. The goal of this paper is to create a tool to "browse" audio efficiently, by using the human ability to spatially access information.

The "cocktail party effect"[1] is the foundation on which this paper relies; we have the ability to selectively attend to one sound source in the presence of other sounds and background noise, and the ability to listen to a background channel. By taking advantage of the cocktail party effect, and by introducing the idea of spatial mapping of audio, we created an audio browsing environment in which one can browse audio data by switching focus between multiple sound sources, and can use spatial memory to access information.

This paper first gives an overview of other approaches used for audio browsing, and then introduces the key ideas of this paper: the simultaneous presentation of audio and the spatial mapping of audio. The iterative design approach toward implementation is reported next; the audio presentation is designed to help form spatial memory and to enhance selective listening, and new interaction methods are developed to reduce errors and enable fine grain control of audio playback.

RELATED WORK

Audio Notebook

Audio Notebook [14] is an enhanced paper notebook, which allows a user to capture and access an audio recording of a lecture or meeting in conjunction with notes written on paper. Users can access the recorded audio by flipping pages or by pointing to a location in the notes. With the Audio Notebook, users often remember the mapping of physical location in their notes to the desired audio information.

Filochat

Filochat [16] co-indexes speech recording to pen-strokes of handwritten notes taken with a digital notebook. Users can access a portion of audio data that is associated with a specific note by gesturing at the note. Filochat allows users to access audio data directly by using spatial memory of written notes.

Unlike Audio Notebook or Filochat, this paper focuses on the creation of audio-only browsing environments. Taking advantage of the omni-present and omni-directional nature of our hearing, this paper implements a system that utilizes hearing as another channel of input which is available for listening even when you are busy writing or driving. Audio Notebook and Filochat are relevant to this work since they utilize spatial memory to access audio recordings. While they have visual marks on notes to help remember location, this paper takes on the more challenging task of using spatial memory in an audio only environment.

Applications of spatialized audio

Cohen and Ludwig [4] used acoustic processing to give audio "windows" status as foreground/background auditory prominence. The virtual meeting room developed at AT&T Bell laboratories [13] used spatialized audio to provide information about the connectivity, presence, focus and activity of the participants using sampled sounds such as keyboard clicks as well as speech. Pitt and Edwards [9] examined the use of stereo audio for manipulating objects through mouse movements for blind users. Objects in the space make sounds simultaneously, and the intensity of the sounds reflects the distance to the cursor. Users can

[1] This research was completed at the Media Laboratory while the first author was a student there.

approach an object by operating the mouse so the sound of the desired object becomes louder. Hudson and Smith [5] is a recent example of a long history of using sound to represent or summarize complex data, but our current work focuses on the audio recording itself as the data.

SpeechSkimmer

SpeechSkimmer [2] provides a user interface for skimming or browsing speech recording. It automatically selects and presents salient audio segments as well as reducing playback time by time-compression and pause removal. It provides a browsing environment but does not make use of spatialized audio or simultaneous presentation.

AudioStreamer

AudioStreamer [11] creates an audio-only browsing environment that enhances the listener's ability to browse audio data by taking advantage of the "cocktail party effect." It presents three audio data streams at three distinct fixed locations in a virtual audio space. The spatial arrangement of the three sound sources facilitates the separation of the simultaneously presented multiple audio data, and allows users to attend to one of them selectively. AudioStreamer enhances our ability to selectively attend to the source of greatest interest by making it acoustically prominent when the user leans toward one particular sound source. It also augments our ability to perceive events in the nonattended audio channels by auditorially alerting users to salient events on those channels. Users of AudioStreamer can use their spatial memory for audio navigation, such as a topic which was heard on the "left channel."

AudioStreamer showed the potential for simultaneous listening for audio browsing. Motivated by AudioStreamer, this paper implements an alternative form of spatialized simultaneous listening for more efficient temporal navigation of a single audio recording. The major differences between AudioStreamer and this paper are (1) the number of audio data streams, and (2) the location of sound source. By playing a single audio stream by means of moving sound sources, the system of this paper maps time to space, while AudioStreamer maps three audio streams to three locations by playing three audio streams through three fixed location sound sources.

DYNAMIC SOUNDSCAPE

This section describes Dynamic Soundscape, the browsing system of this paper. This section first describes what users hear, and then introduces two key ideas on which the system is based: simultaneous presentation of audio and spatial mapping of audio.

Basic idea of the browsing system

Figure 1 illustrates the concept of the auditory space created by the system. The objects in the figure are invisible audio objects.

The basic object of the Soundscape is a "Speaker," a moving sound source orbiting the listener's head. There can be multiple Speakers simultaneously playing different portions of the same audio data stream, although there is one Speaker when the system starts playing. When the system starts, a Speaker is created at some point in an orbit around the user. The Speaker orbits the user's head as it plays the audio data, and so creates a map between time and space.

When the user wants to re-play a topic that he/she heard, he/she indicates the position where the topic was presented by pointing in that direction. Another Speaker is created at the point and begins playing from the point in time of audio data presented there (Figure 2). The original Speaker continues playing after the new Speaker is created, so the user hears multiple portions of the recording simultaneously from the Speakers. The original Speaker decreases its loudness until the user returns his/her attention to it by indicating the current position of the original Speaker.

The user can jump ahead by indicating the appropriate position ahead of the original Speaker. A new Speaker is created there, playing the appropriate audio data for that position in the sound file, and the original Speaker again continues to play and decrease loudness. Though the user jumps ahead, he/she can hear the skipped audio data from the original Speaker which is running after the new one, and when he/she finds something interesting from the original Speaker, he/she can switch back to the original one by indicating it.

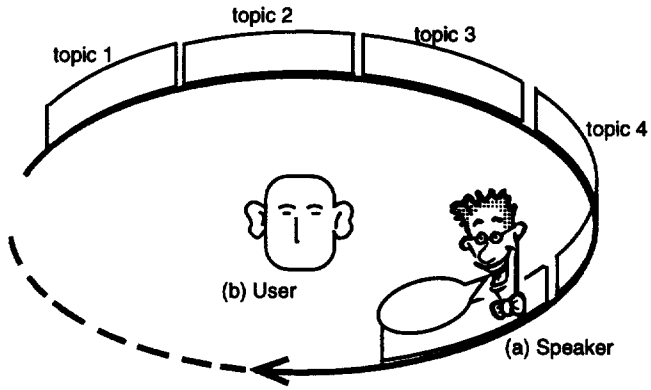

Figure 1: The concept of the auditory space created by the system. A Speaker (a) in the virtual acoustic space speaks audio data as it goes around the user (b).

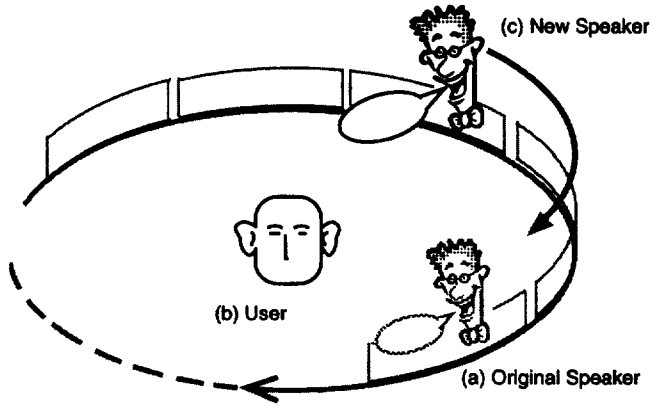

Figure 2: Upon user's request, a new moving *Speaker* (c) is created where the user points. The original *Speaker* keeps going. The user hears multiple portions of the audio stream simultaneously.

Simultaneous presentation of audio

Simultaneous presentation of multiple portions of a single audio recording is one of the key ideas of this paper. Instead of fast-forwarding or rewinding, the user can browse the audio data by switching attention between multiple sound sources, each of which plays different portions of the same recording, analogous to visual browsing in that we move our focus around the document. Even when the user is concentrating on one sound source, he/she can hear other portions of the audio data from other sound sources in the background, and he/she can switch to another sound source upon finding a more interesting topic in it.

Spatial presentation of audio

Spatial presentation of audio data is the other key idea of this paper. In contrast to the lack of temporal attributes, the visual attribute of spatial location is commonly and automatically encoded to the memory of events. Whether it has a real or imagined context of space, it is frequently recalled and intimately associated with the recognition of the events, and enhances the memory of the event [12] [8].

The spatial presentation proposed in this paper allows the use of spatial attributes of our memory to compensate for the weakness of temporal recall. By associating the temporal axis of audio data onto the spatial location in virtual acoustic space, the browsing environment will enable users to navigate through the audio data based on their spatial memory, and by means of spatial expression of amounts such as distance or angle. Instead of using temporal expressions such as "20 seconds ago" or "20 seconds later," users can access audio data spatially as "the news I heard when the speaker was at my back-left" or "supposed to appear around here."

INITIAL IMPLEMENTATION

To implement the browsing system described in the previous section, we took an iterative design approach; we started with a rough implementation of the idea, and evolved the system based on feedback of the implemented system. This section describes the initial implementation of the iterative design, and the problems found.

Speaker motion

We initially tested three models of Speaker motion: (a) mono-directional straight line, (b) bi-directional straight line, and (c) a circular orbit. The orbital motion was chosen because (1) it can present a long audio recording continuously, (2) it provides a two dimensional space for accessing Speakers, and (3) the disadvantage for users who cannot tell front from back can be reduced by suggesting that the Speakers are moving clockwise[2].

Interaction

In the initial implementation, the simple action of "pointing" is the only way to interact with the system. Users can "point" to a location on the round path by pushing on the touchpad interface (see Figure 4). When the user points to a location where an active Speaker exists, the input is interpreted as the "*switch focus*" command resulting the change of focused Speaker which is presented most loudly. When the user pointed to a location where no Speakers exist, the user's input is interpreted as the "*create new Speaker*" command which creates a new Speaker to play the audio recording from the point that corresponds to the location. There can be at most four Speakers at the same time. When the user requests a fifth Speaker, the least focused Speaker is terminated, and used to create the new Speaker. By creating Speakers, users can play the desired audio based on their spatial memory, and by switching focus between multiple Speakers, users can browse the audio recording having multiple view windows provided by the simultaneous presentation of single audio recording.

System architecture

The hardware configuration of the system is shown in Figure 3. The Sparcstation and the storage device work as the audio server, which plays multiple digitized audio signals. The Sparcstation has two stereo ports, so the system can have at most 4 monaural outputs. The Network Audio Server (NAS), a separate process from the application program, interleaves the left and right channels of audio from separate files into stereo digital audio streams.

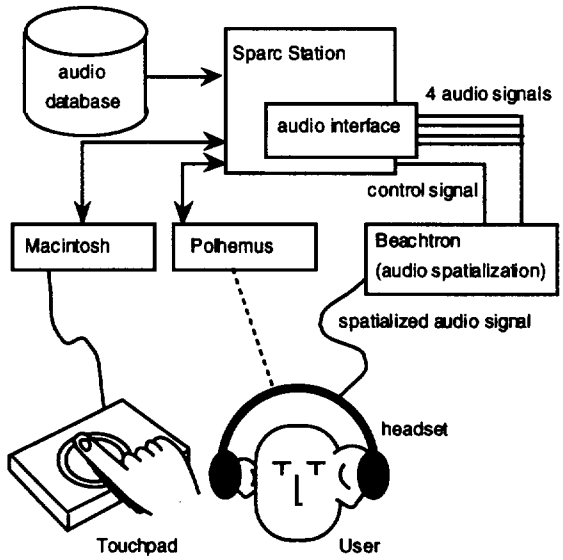

Figure 3: System configuration

Two Crystal River Beachtron audio cards, installed in the audio spatialization server PC, receive the four audio signals from the Sparcstation. The Beachtron cards spatialize the audio signals, locating monaural audio signals at the specified location in virtual acoustic space.

The Polhemus position sensor is mounted on the headset, and measures the location and direction of the user's head. For locating sound, we naturally use the cue of the change of audio stimuli as we move our heads. By actively changing the virtual audio environment according to the motion of listener's head, the sense of space of the virtual audio can be enhanced [7]. The information captured by the Polhemus sensors is sent to the audio spatialization server, so it can change the relative position of the sound sources.

[2] When a Speaker is moving right, it should be in front, while when moving left, it should be in back.

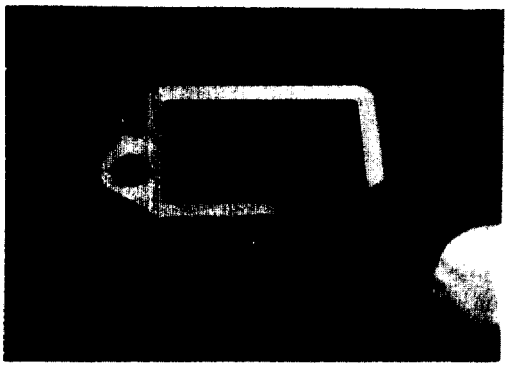

Figure 4: Touchpad interface. The template is attached to the surface of the touchpad, so the user can feel the shape of round path without seeing the device

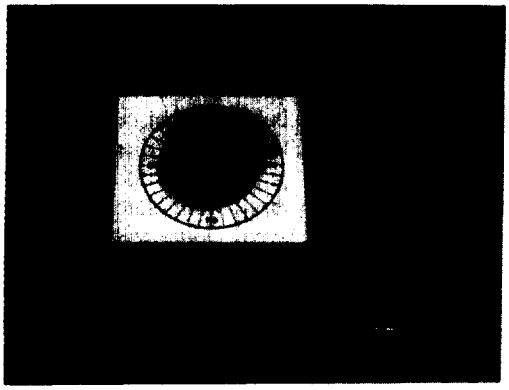

Figure 5: Knob interface.

The Macintosh computer receives the user's input through interface devices connected to the ADB bus, to which various devices are available in the market. In this initial implementation, the touchpad interface (Figure 4) and the knob interface (Figure 5) are connected. The Macintosh is connected to the Sparcstation via the serial interface.

Feedback from the initial implementation

Evaluation of the first system with the touchpad interface revealed several problems.

Problem I: Difficulties in remembering topic locations

Memory of the spatial locations of audio events is one of the key issues of this paper. By playing an audio recording through moving Speakers whose location is a function of time, a mapping between spatial location and time is formed, and users can employ spatial location to navigate through the temporal audio. We expected that users could remember the location of topics (such as a news about "election") or events (such as a sound of someone shouting) in the audio recording; such memories of location are essential to use the spatial location for audio navigation.

However, in experimental use of the browsing system in which Speakers move at the speed of 6 degrees per second, it seemed hard for users to remember the locations of topics and events in the audio recording. Positions seemed to become vague because the Speaker moved while they were being presented.

One cause of this vagueness of memory about the location may be the motion of the sound sources. If our listening ability itself is different when the sound sources are moving from when the sound sources stay still, as our sight of moving objects is different from the sight of still objects, the basic idea of the browsing system must be reconsidered.

Another cause may be that the speed of the Speaker was inappropriate. The speed of 6 degrees per second was chosen because at this speed the Speakers seem to move, and there is enough spatial resolution to access multiple events within a topic. However, it seemed too fast to remember the location of events and topics.

Further evaluation of the manner and the speed of Speakers' motion seemed to be essential to make the system usable.

Problem II: Difficulties in selectively listening to virtually spatialized audio

Although we have the ability to listen to a sound selectively, when adding the Speakers in order to play other portions of audio, it becomes hard to hear one sound selectively. Selective listening seemed to be harder in the virtual audio space than in the natural audio space because of the less than perfect spatial localization. Another factor is that the multiple Speakers may be playing the same talker's voice, since they all play the same recording. Difference of voice, which is one of the factors that contribute to the ability of selective listening [3], is small. A study about the way we selectively listen to one sound source among multiple sounds was needed to provide the basis to build a human interface to enhance the selective listening in the virtual audio space.

Problem III: Error in locating the sound source

For users who do not perceive the audio spatially, it is impossible to remember the spatial location of topics. Even for users who perceive the audio spatially, but imperfectly, the error in locating sound sources results in memory of the wrong location. Since all users use a common HRTF (head related transfer function), which should vary slightly between each user because each of us has differently shaped ears, there always is a gap between the intended and perceived position of the sound. It is necessary to bridge that gap.

Problem IV: Resolution of memory of sound location

Our memory about the location of topics in the audio recording has poor resolution. We usually memorize the location in quadrants or 12ths of a circle, such as saying "a topic in left front," but never say "the topic at 38 degrees from the front." When pointing to a location to access the audio corresponding the location, we may be able to point close to the correct location, but it is almost impossible to pinpoint the position. It is necessary to estimate the probable position that the user might desire, and to be able to adjust after hearing an incorrect audio selection.

Problem V: Indirect pointing interface

Errors also occur when pointing to a location by using the pointing device. Even if the user has an accurate memory of location of the sound sources, an error may occur when he/she transfers the location in the space of memory, which is ideally same as the space of audio, onto the location in the space of interface device. In the case of the touchpad

interface of this initial implementation, the users have to transfer the location on the 40 inch radius circle around their heads onto the location on the 1 inch radius circle on the device. A direct means of pointing to the location where he/she hears the audio is necessary to reduce this error.

ITERATIVE DESIGN: AUDIO PRESENTATION

Presentation of audio was redesigned to solve the first two problems of the initial system. The mapping of time to space has been redesigned to solve Problem I: Difficulties in remembering topic locations, and the slight head motion was utilized to solve Problem II: Difficulties in selectively listening to virtually spatialized audio, enhancing the human ability of selective listening.

Mapping time to space

Using the initial system, it seemed difficult to remember efficiently the topics and their locations. Two factors might contribute to the difficulties: the motion itself, and the inadequate speed of motion. If the motion itself is the problem, discrete motion, in which Speakers move once in a while, should work better. If speed is the problem, slower speed lets users memorize better. We examined these two approaches, discrete motion and slower motion, by comparing three types of motions:

(a) original fast continuous motion, in which Speakers move at 4.8 degrees per second,

(b) fast discrete motion, in which Speakers move once in approximately five seconds, at the rate of 4.8 degrees per second, and

(c) slow continuous motion, in which Speakers move at 1.2 degrees per second.

Four subjects were asked to listen to a 5 minute recording of a radio news program being played through a Speaker that moves in one of the three motions. Each subject performed three sessions with three motions in a random order. After each session, subjects were asked to list all the topics they remembered, and the location of the topics.

Result

With the slow continuous motion (c), all subjects remembered more topics and their locations than with other motions. Even subjects who did the session with slower motion first and the session with fast motion next remembered the location of topics better in the slower motion. A subject who tried hard to remember the location of topics could tell locations of topics which sometimes span 180 degrees, but only about the topics presented at the beginning and the end of the session, reflecting the characteristics of short term memory. The slow continuous motion was also the motion most preferred.

Discrete motion (b) did not make for better memory. Furthermore, the sudden jump of the discrete motion sometimes made it difficult for users to follow the Speaker, especially in multi Speaker situations.

Along with asking what they remembered, subjects were asked about the resolutions with which they localized audio events. Though dependent on how well audio localization worked, most of the subjects answered it was between quadrants and 12th of a circle, which is much more than left and right, but much less than 10 degrees. The ordinary length of single topic in the news program used in this experiment was 30 seconds. With the slow continuous motion (c), the length corresponds to 36 degrees, which is close to the resolution of the memory of location of audio events that the subjects reported.

The slow continuous motion (c) was chosen as the motion of Speakers of the browsing system.

Enhancing selective listening

In the initial implementation, it was hard to listen to one of the multiple sound sources, though we have the ability to listen selectively in the natural environment. This is certainly in part due to the fact that each Speaker may talking in the same announcer's voice. To enhance the ability of selective listening in the virtual acoustic space, we first observed our behavior while listening to a sound, and then developed an interface based on natural behavior.

Observation

In order to find the natural behavior that could be used as a clue for the system to know the user's interest, we did a brief observational study of listening to a sound in the presence of background noise. To simulate the situation of the system, we placed three[3] loudspeakers around the subjects, and played three audio streams of conversation. Subjects were asked to listen to one sound source and understand the contents. The behavior was video taped by a camera in front of the subjects. Some subjects participated in an additional subsequent experiment, in which binaural virtual audio was used instead of the three loudspeakers.

Similar experiments have been done to observe the head motion when locating a sound source in space [15] [6]. Strategic motions to find the sound sources were observed in those experiments. This experiment focused on listener motions for selective listening, for better reception of the desired sound.

In the first experiment, some subjects moved their bodies radically, and others did not. Selective listening is performed by the combination of both physical movement and internal filtering within our brain. Some subjects moved their bodies actively to hear better, and others could listen selectively without moving.

Slight adjusting head motions were common even among the subjects who did not move much. They adjusted their head location and direction repeatedly to find the best location and direction. Leaning toward the speaker was often observed in the adjusting motion (Figure 6). Though leaning was not always directly toward the sound source, it was close to the direction of the sound source.

In the experiment with virtual audio, the subjects tended to not move their heads. If the audio spatialization worked properly for the subjects, adjusting their head location should be beneficial for getting better reception of the desired audio. After the experiments, subjects said that they did not

[3] The hardware supported 4 output channels, and we reserved one for the audiocursor, which will be described in the next section, leaving three max for Speakers.

move their heads because they knew the audio was presented through the headsets, and thought that adjusting head position was not beneficial. Whether the audio spatialization worked well or not, the change of audio in response to the subjects' head motion was not clear enough to be beneficial.

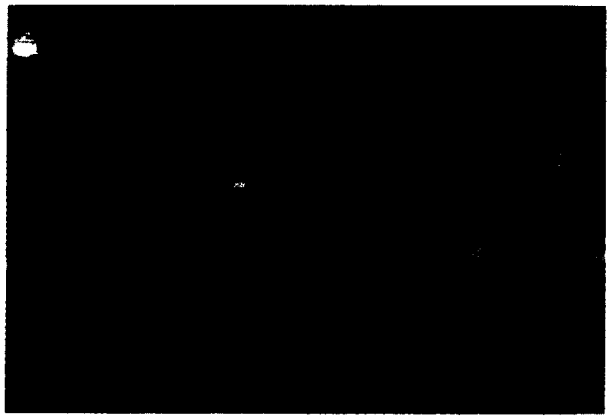

Figure 6: Leaning head toward the speaker on the left

Design of head interface to enhance selective listening

We learned from these observations that people move their heads to help selectively listen. This cue was added to the system, using head motion to enhance the human ability of selective listening. The system measures the direction of leaning with the Polhemus sensor attached to the headset. The system changes the loudness of each Speaker according to the angular distance between the angle of the Speaker's location and the angle of the direction of leaning. The change of the loudness is proportional to the angular distance. So the closer the user leans his/her head toward a Speaker, the louder the Speaker plays.

Since the change of loudness is a continuous function of the angular distance between the direction of leaning and the location of the Speaker, users can adjust their leaning direction by responding to the feedback from the system. Such adjusting motion induced by the feedback from the system is similar to the natural motion while listening.

The system makes a Speaker at most 8 dB louder than others when it is close to the direction of leaning, an exaggerated feedback, which never happens in the natural environment. This exaggerated feedback makes the benefit of leaning motion clear to the user, artificially enhances selective listening.

ITERATIVE DESIGN: INTERACTIVE ACCESS TO AUDIO

It was difficult for users of the initial system to point to the correct location to access the desired portion of audio. There were three types of obstacles: errors in locating sound (Problem III), insufficient resolution of memory of sound location (Problem IV), and errors in pointing (Problem V). To improve the accuracy and the resolution to access the desired data, three new interaction methods were developed.

Grab-and-move interaction

In the initial system, it was difficult for users to point to the correct location of the desired audio because their memories of location of audio events have inadequate resolution. The slower speed motion, which was chosen as described in the previous section, maps longer portions of the audio recording to a unit space; as the result the resolution of pointing decreases. In order to enable fine grain control of audio, the system employs a "grab-and-move" interface, with which users can adjust the location interactively after hearing the audio which is of the wrong location.

Like the "pointing" interface of the initial implementation, users request the system to play a portion of audio by pointing to the location that corresponds to the audio. When there is a Speaker at the location, the system puts the Speaker under the user's control, which is the "grabbed" state. If the audio that the grabbed Speaker begins playing is different from what he/she expected, the user can move the grabbed Speaker to adjust the point to play. While the Speaker is moved, it plays small segments of the audio of the location, so the user hear the contents of the audio.

When there is no Speaker at the location the user pointed to, the system creates a new Speaker there, and starts playing the audio recording from the corresponding point. The system has a table of times which are probable boundaries of topics. The preprocessor, which was developed for Newscomm [10], generates the table based on acoustic cues such as the long pause, change of talker, or emphasized voice. When the system decides the position to play from, it chooses a time closest to the pointed location from the boundary table. This is to enable more efficient navigation by guessing a more salient point from which to play.

While a Speaker is grabbed, it is played louder than others to notify the user it is grabbed. After 3 seconds without moving, or by the input from the interface devices, the grabbed Speaker is "un-grabbed" and returns to normal.

Audio cursor

For most users testing the browsing system, audio spatialization was less than perfect. There is always a mismatch between the locations of the sounds that the users perceive and those at which the system intends them to be. To enable precise interaction with the objects in the virtual audio space, a means to bridge the gap is necessary.

The "audio cursor" is an audio object in the virtual audio space of the system. It continuously plays the sound of a vibrating spring in a tube (zubetube.au), a noise with a distinctive rhythm, while it is turned on by the user. It provides location feedback, so the audio cursor moves within the virtual audio space as the user operates the interface device. Before "grabbing" an audio object, the user moves the audio cursor to the location of the audio object, and acoustically overlays the audio cursor on the audio object.

The touchpad interface and the knob interface were modified to adopt the grab-and-move interaction and audio cursor. The user can move the audio cursor by moving his/her finger on the touchpad or by rotating the knob, grab a Speaker by pressing the touchpad/knob, and move the Speaker by moving his/her finger pressing down the touchpad surface or by rotating the knob pressing down it.

Point-by-hand interface

A direct means of interaction is needed in order to reduce the errors that occur in transferring the location in the virtual audio space to the location in the space of interface device. The point-by-hand interface is a hand gesture interface, with which users can access their desired data by directly pointing to the location where they hear the audio.

The interface device to detect the hand gesture is built with the Fish sensor, which is a non-contact sensor based on the interaction of a person with an electric field [18]. As shown in Figure 7, with a transmitter on the chair on which the user sits, and four receivers hang over the user, the Fish can detect the distance between each sensor and the user's hand as the intensity of electric field. After a calibration session, which is necessary to adapt to the various sizes and shapes of user's body, the system can compute the x-y coordinates and the height of user's hand.

With the point-by-hand interface, the user turns on the audio cursor by raising a hand, and moves the audio cursor by moving the hand. To grab a Speaker, the user moves the audio cursor to the location of the Speaker, and stretches the arm like grabbing an apple on a branch. The grabbed Speaker is kept grabbed until he/she lowers the hand.

Figure 7: The "point-by-hand" interface in use. Four metal balls are the receivers. The transmitter is on the chair.

DISCUSSION

User feedback

Memory of spatially mapped audio

In contrast to the initial implementation where users could not remember the location of audio events, most users reported that they could use their spatial memory for audio navigation with the refined system. When the Speakers were moving at an adequate speed to form memory of the topics, the space seemed to help users to memorize the topics. By observing subjects, we are led to believe that the association between the topics and spatial locations helps to transfer the memory of topics to the long term memory. The spatial presentation provides an environment to organize information, with the mapping which associates the contents of the topics to spatial locations. This association aids recall of the story topics. This memory effect similar to the "Simonides memory palace"[17] is made available in the audio only environment by this system.

Head interface: enhancement of selective listening

The head interface implemented in the system enhances the human ability of selective listening based on the leaning behavior which was often observed in selective listening in the natural environment. It imitates the interactive process between listener and the audio space; as the listener moves his/her head, the reception of the sound source changes, and then the listener adjusts his/her head repeatedly to get better reception. For many users, it was a natural iterative process, and they could comfortably use the interface to listen selectively. With this interface, users can naturally and quickly switch their attention among multiple sound sources. This allows browsing by switching attention, instead of fast-forwarding and rewinding.

Interface design: large vs. small interface

The large scale "point-by-hand" interface and the small scale touch pad interface were compared by several users from the point of accuracy, and ease of use. They reported that both interfaces were easy and accurate for navigating audio. The large "point-by-hand" interface was preferred because it had the scale closer to the scale of the path of the Speaker.

We expected that the small interface would place higher cognitive load on users because of difficulties with the cross-space mapping between the large virtual audio space and the small interface space. However, most users did not find it hard to use the small interface device, because they were familiar with controlling small devices such as a mouse.

To use the "point-by-hand" interface, users had to learn the height to raise their hands to control the audio cursor or to grab a Speaker. However, for those who were used to the operation of the interface, it was an easy and direct interface.

Audio cursor

The audio cursor was helpful for users for whom the audio spatialization did not work well. By moving the audio cursor, they could learn the correlation between locations in the virtual audio space and locations in the space of interface devices. With the "point-by-hand" interface, the audio cursor sounds at the location close to the user's hand. It produces an illusion of moving the audio cursor by their hand, and enhances the sense of space of the virtual audio space.

Future directions

Adaptive mapping

The speed of Speakers was chosen based on the typical length of topics in the audio recording used in the experiments. The typical length of topics may differ by the type of the audio recording, such as radio news, recordings of lectures, or audio books of novels. It is desirable to change the speed of Speakers based on the type of the audio recording, or to develop more adaptive mapping that maps each topic in an audio recording to the same amount of space by changing the speed according to the length of the topic.

Enhancing selective attention

Some users reported that it was difficult to notice salient topics spoken by a non-attended Speaker. In this system, all Speakers are presented at the same loudness unless the user leans toward a Speaker. Users tend to move their heads and switch around the Speakers once in a while. The head interface, which enables easy quick switching between the Speakers, allowed such hopping around activity which is analogous to the eye movement in browsing printed documents. Although they could patrol other Speakers by hopping around, users sometimes miss interesting events in the background channel because of the temporal nature of audio. Approaches developed in AudioStreamer [11], which arouse the user's attention at prominent events, should be combined with this browsing system.

CONCLUSION

This paper presented an iterative design approach toward a browsing environment that provides users with a spatial interface to access temporal audio data. By mapping time to space, and simultaneously presenting multiple portions of an audio recording, the system enables users to browse audio data spatially. Although further work to implement adaptive mapping is necessary, this paper suggested the approximate guideline of mapping that a topic should be mapped to a unit area of our memory of sound location, which is generally a quadrant or a 12th of a circle.

This paper also covered methods of interactive access to the system. The grab-and-move interface enables fine grain control of audio, and compensates for the small spatial resolution of our memory of sound locations. The audio cursor compensates for localization error, which largely depends on the individual listening characteristics, and it also enables precise access of audio objects by acoustic overlay of the cursor and the object. The point-by-hand interface enables direct and intuitive access to the audio objects. Along with the audio cursor, the "point-by-hand" interface creates the feeling that the user is touching the audio object, and increases the spatial experience of the virtual audio space. The head interface contributes to enhance the ability of selective listening in the virtual audio space.

The implemented system proved that spatial memory of audio events is usable for audio browsing. We expect spatial mapping of audio data for temporal navigation of audio to be a new dimension of application of the audio spatialization technologies.

ACKNOWLEDGMENTS

We thank Jordan Slott who implemented the Network Audio Server (NAS), and Atty Mullins who implemented the AudioStreamer which gave us the motivation and technical basis of this work. We also thank several research associates of the MIT Media Lab for testing the system, and the anonymous reviewers for their constructive critiques.

REFERENCES

1. Arons, B., A Review of the Cocktail Party Effect. *Journal of the American Voice I/O Society*. 1992.
2. Arons, B., SpeechSkimmer: Interactively Skimming Recorded Speech, *procs. of UIST '93*, 1993.
3. Cherry, E. C., Some experiments on the recognition of speech, with one and two ears, *Journal of the Acoustic Society of America*, Volume 25 1953.
4. Cohen, M. and Ludwig, F. L., Multidimensional Audio Window Management, *International Journal of Man-Machine Studies*, Vol. 34, Academic Press. 1991
5. Hudson, S. E. and Smith, I., Electronic Mail Previews Using Non-Speech Audio, *procs. of CHI 96*, ACM. 1996.
6. King, W. J. and Weghorst, S. J., Ear Tracking: Visualizing Auditory Localization Strategies, *procs. of CHI 95*. ACM. 1995.
7. Loomis, J. M., Hebert, C., Chcinelli, J. G., Active localization of virtual sounds, *Journal of Acoustical Society of America*. 1990.
8. Mandler, J. M, Seegmiller, D. and Day, J., On the coding of spatial information, *Memory & Cognition* 1977, Vol. 5. 1977.
9. Pitt, I. J. and Edwards, A. D. N., Pointing in an Auditory Interface for Blind Users, *procs. of the 1995 IEEE International Conference on Systems, Man and Cybernetics*, IEEE. 1995
10. Roy, D. K., NewsComm: A Hand-Held Device for Interactive Access to Structured Audio, *procs. of CHI 96*. ACM. 1996.
11. Schmandt, C., Mullins, A., AudioStreamer: Exploiting Simultaneity for Listening, *procs. of CHI 95*. ACM. 1995.
12. Schulman, A. I., Recognition memory and the recall of spatial location, *Memory & Cognition* 1973, Vol. 1, No. 3.
13. Seligmann, D. D., Mercuri, R. T. and Edmark, J. T., Providing Assurances in a Multimedia Interactive Environment, *procs. of CHI 95*, ACM. 1995
14. Stifelman, L. J., Augmenting Real-World Objects: A Paper-Based Audio Notebook, *a short paper in CHI 96*, ACM. 1996.
15. Thurlow W. R., Mangels, J. W., and Runge, P. S., Head Movements During Sound Localization, *Journal of the Acoustical Society of America*, 1967.
16. Whittaker, S., Hyland, P. and Wiley M., Filochat: Handwritten Notes Provide Access To Recorded Conversations, *procs. of CHI 94*, ACM. 1994.
17. Yates, F. A., *The Art of Memory*, Routledge & Kegan Paul. 1966.
18. Zimmerman, T. G., Smith, J. R., Paradiso, J. A., Allport, D. and Gershenfeld, N., Applying Electric Field Sensing to Human-Computer Interfaces, *procs. of CHI 95*, ACM. 1995.