INTERACTIVE LISTENING TO STRUCTURED SPEECH CONTENT ON THE INTERNET

Makoto J. Hirayama, Taro Sugahara, Zhiyong Peng, and Junichi Yamazaki

Hewlett-Packard Laboratories Japan

ABSTRACT

Interactive information browsing of World Wide Web is a key application of the Internet and visual web browsers are widely used to access information. However, visual web browsing is not suitable in some circumstances such as in mobile environment. Therefore, we propose interactive listening to structured speech content for accessing information. Our proposed model of the interactive information listening services is that structured audio contents (HyperAudio) in a HyperAudio server are listened using a HyperAudio player whose appearance is similar to a portable radio. Unlike radio broadcasting programs, the HyperAudio contents have logical structures and hyperlinks so that listeners can listen desired information interactively. To put such logical structures into audio, a simple markup language was used. A prototype system of the HyperAudio server and players was implemented to test and evaluate feasibility and usability of the HyperAudio architecture.

1. INTRODUCTION

Auditory user interfaces will be important especially for mobile information appliances such as cellular phone based personal digital assistances (PDA) or enhanced car navigation systems. For such devices, one of the key applications is information browsing. Interactive information browsing of World Wide Web with a visual web browser is very useful, because it has structured representation of text and hyperlink navigation written in the HyperText Markup Language (HTML) [1][2][3]. However, in visual browsing there are two main drawbacks, 1) human vision is occupied therefore it cannot be used during walking or driving, and 2) a large screen is needed therefore visual devices are not portable. Our challenge is to enable interactive information listening to speech content with user audio player agent devices.

2. INTERACTIVE LISTENING

2.1. Advantages of Audio

In some circumstances, auditory human interfaces [4] are more useful than visual ones. We think that following characteristics of auditory human interfaces are important for accessing information by audio:

1. Human vision is not occupied. Thus, it is possible to do other things simultaneously. For example, listening during walking, driving a car, standing in a train, or washing dishes in a kitchen is possible.
2. Devices can be small in size and light in weight. Large screens for visual representation are not needed. Therefore, a pocket sized device which is suitable for mobile circumstances can be used.
3. Listening is easier than reading a lot of texts on screen, especially in a casual and relaxed situations in a living room.
4. A telephone can be used as a user interface device. Telephones are very popular and widely spread devices all over the world. Accessing information via a telephone is easy for all generations all over the world.

By these characteristics, we consider that auditory human interfaces for information accesses are promising. However, to use auditory human interfaces for information accesses, we think that we must solve one key issue. That is how to design convenient and comfortable auditory human interfaces for users to reach and get desired information effectively. In the next section, our proposed concept for interactive information listening will be presented.

2.2. Interactive Listening

Our proposed model of the interactive information listening services is that structured audio (i. e., non-sequential audio, we call it HyperAudio) in a HyperAudio server are listened using a HyperAudio player whose appearance is similar to a portable radio. Unlike radio broadcasting programs, the HyperAudio contents have logical structures and hyperlinks so that listeners can listen desired information interactively.

For example, in the case of news contents, the first audio file contains all of today's headlines and the second file contains full stories for the corresponding headlines. Each headline and a corresponding full story is hyperlinked. When a user accesses the first file, he/she will listen to today's headlines. During playback of the headlines, if the user wants to know more details for a specific headline, he/she pushes the "Jump" button on the player device, then the player plays back a specific portion of the second file so that he/she can listen to a desired full story selectively. Figure 1 shows the structure of this example. This auditory human interface is more interactive than those in radio broadcasting programs (user interactions for radio broadcasting are only power on/off and program selection

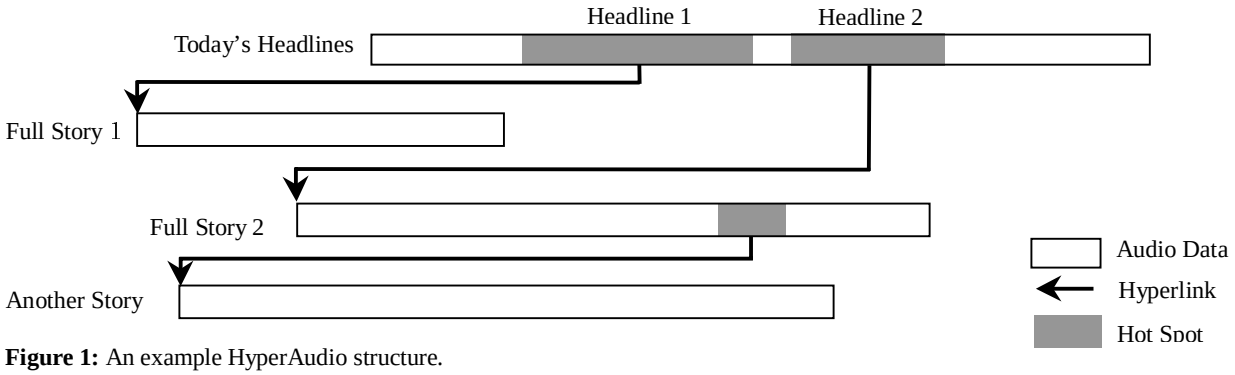

Figure 1: An example HyperAudio structure.

is by tuning). Thus, we call this scheme as the interactive listening.

2.3. The Reason for Structures in Audio

For information listening, radio broadcasting is very popular. We can listen much information from radio. But, in radio broadcasting programs, there is no real-time interaction between contents and listeners except for selecting a program by tuning. Therefore, listeners must listen an entire program from start to end, even if some parts of the program are not favorite ones. For example, in a weather report, listeners usually want to listen a weather report for only an area which they live in. But, in an actual program, a weather report for all area in the country must be listened to get one for a desired area. Or, in a traffic report, they want to listen a traffic report for only roads which they are driving on. Or, in a distance learning program, students may want to repeat a specific part or listen to more detailed explanation. If the radio is interactive, we can get desired information more effectively. That is why we put logical structures and hyperlinks onto prerecorded audio files of speech contents to be listened interactively.

Our proposed structure is similar to existing visual web pages. You may think that listening to existing visual web page on the Internet through a text-to-speech synthesizer is a good way to get information interactively by listening. Such browsers are known as web readers. Emacspeak [4], Home Page Reader [5], etc. are examples of web readers. This approach is useful to get information from existing web pages which have been already made. However, almost all of the existing web pages have been designed assuming that users does not listen to the contents but view them. We tried to use a web reader, however, it was very hard to reach and get information. The first reason for making listening impossible, existing web pages have so many visual elements that we cannot understand these pages only from text elements. Second, some verbal expressions for viewing are not matched to ones for listening. As a simple example, "Click Here" for viewing should be replaced by "Push Now" for listening. Third, we felt that effective textual presentations for viewing may be different from effective aural presentations for listening. Furthermore, logical structures which can be easily understood by viewing may not be understood by listening.

If authors wrote contents very carefully by imagining that they may be not only viewed but also listened, it may be possible to share texts both for viewing and listening. For a such purpose, aural cascading style sheet (ACSS) [6] is available now. But, we feel that it is not easy to write contents comfortable enough both for viewing and listening. We concluded that separate contents for listening from ones for viewing should be designed independently. Even if source information itself is same, the most appropriate presentation is different depending on methods of access. Thus, we proposed the approach of putting structures on prerecorded audio files.

2.4. An Interaction Model

In traditional interactive systems such as interactive voice responses (IVRs), the interaction model is like a dialogue between a human operator and a user. Especially in systems that use automatic speech recognition and text-to-speech synthesizer, a dialogue model that emulates conversation between humans is used. Usually, question and answer type navigation is done to reach and get information or to do tasks like airline ticket reservations.

On the contrary, our proposed interaction model in this work is somewhat different. In a typical case, once a user reached a desired program through menu style hyperlinks on a top page and some of subsequent pages, he/she will listen to a content without interaction just like a radio program. Only when he/she wishes to interrupt and change the sequence of the program, he/she pushes "Jump" button or some other buttons like "Prev," "Next." We call this kind of limited interaction model as "oral presentation model" in contrast with "dialogue model." The difference between these two models is that the dialogue model is, per se, symmetrical communication between humans but the oral presentation model is more asymmetrical one.

In an oral presentation, an entire content is prepared and presented according to a presenter's preplanned sequence. But, audiences can interrupt the presentation and say some words such as "Please repeat this explanation again.", "Please skip this part because we have known this topic well.", "Please explain more details on that.", or "I have a question on that.", etc. In the interactive listening system, the first two can be done by

```
<ilsl>
   <head>
      <meta title="HyperAudio Web Interactive News Service"/>
   </head>
   <body>
      <audio src="todaysheadlines.au" abstract="Today's Headlines">
         <anchor href="fullstory1.au$00:00:02.5" begin="7s" end="18s" title="Full Story 1"/>
         <anchor href="fullstory2.ilsl" begin="18s" end="35s" title="Full Story 2"/>
      </audio>
   </body>
</ilsl>
```

Figure 2: An example ILSL document.

"Prev" and "Next" and the latter two can be realized by hyperlinks ("Jump" button).

3. INTERACTIVE LISTENING STRUCTURING LANGUAGE (ILSL)

To specify structures in time in prerecorded speech files, we used a simple markup language. We call it Interactive Listening Structuring Language (ILSL). In earlier versions, we implemented it as an extension to HTML [1]. In the current version, it is defined as an application of eXtensible Markup Language (XML) [7]. Also in the current version, we have slightly changed element and attribute names and syntax to keep consistency with Synchronized Multimedia Integration Language (SMIL) [8]. Thus, the current version of the ILSL is an extension to a subset of SMIL (Figure 2). ILSL file are text files. Actual audio data are stored in different files. ILSL has mainly following functionality:

Logical Structures. Logical structures like headers, paragraphs, and others can be specified. For example, "Next Header" "Next Paragraph" functions can be implemented in HyperAudio players.

Hyperlinks in time domain. Hyperlinks between a specific portion of an audio file to another can be specified. "Jump" which is the essential function of the interactive listening can be realized.

Audio submits. This is optional. This series of elements are used to specify submitting audio from microphone. This is similar to a text submission of HTML forms. This functionality is useful to provide electric commerce services or free keyword searches to find information.

4. HYPERAUDIO WEB

To test and evaluate feasibility and usability of the HyperAudio architecture, we developed a prototype system (HyperAudio Web) and sample contents.

The prototype system has two basic components, a HyperAudio server and HyperSpeech players. They are connected through Internet. Figure 3 shows the HyperAudio system that we developed. A HyperAudio server provides HyperAudio contents, and a player gets and plays HyperAudio contents according to commands issued by users. As a, an interactive radio (Figure 4) whose shape is of mobile appliance was emulated on a PC screen. Also, we developed a HyperAudio Gateway between Internet and a telephone network to use a conventional telephone as a user interface device.

4.1. HyperAudio Server

The HyperAudio server stores HyperAudio contents and sends them to HyperAudio players. It consists of an ILSL server and an audio server.

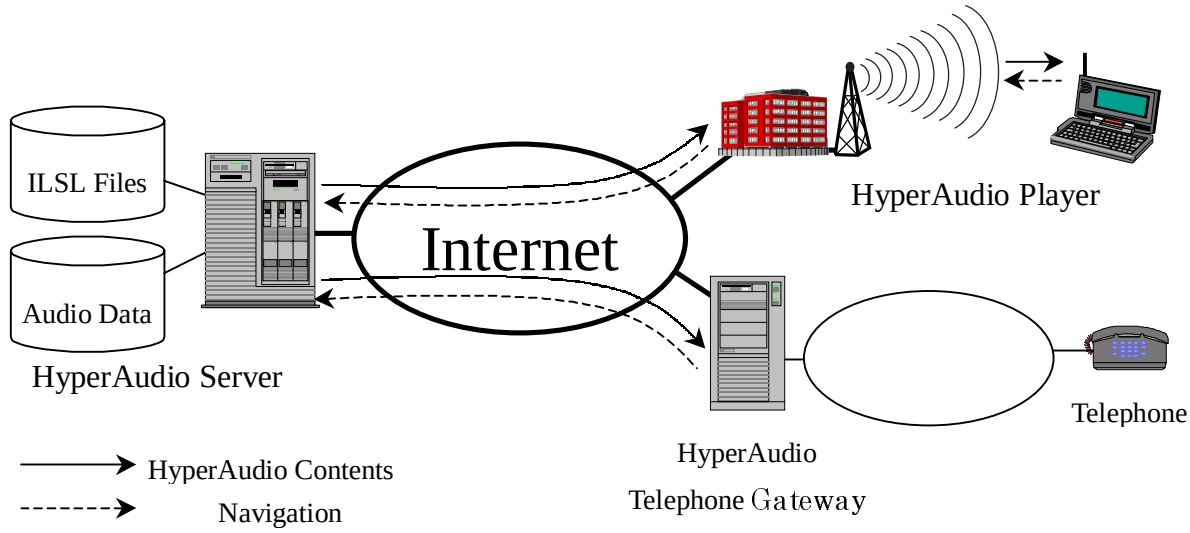

Figure 3: HyperAudio Web System.

Figure 4: HyperAudio Player - Interactive Radio -.

We use a usual web server as an ILSL server. ILSL files are transferred by HyperText Transfer Protocol [9].

An audio server handles audio data. It receives audio data requests from user agents and sends the designated audio data to the player. To transfer audio data from audio servers to user agents, we developed a simple protocol that can transfer an arbitrary part of an audio file by specifying a start time and an end time of the audio data.

4.2. HyperAudio Player : Interactive Radio

An example appearance of a HyperSpeech player – Interactive Radio – is shown in Figure 4. To listen a HyperAudio content, the player gets a ILSL file from the ILSL server and parses ILSL. According to the result of the parsing, it requests audio data specified by an audio file name and start and end time to the audio server. At the same time, in this example implementation, the title of the content and first level header titles in texts are displayed on a small screen. Next, it receives the audio data and plays them. When the user presses "Jump" button, it requests another audio data according to the time of pressing.

4.3. HyperAudio Player : Telephone and HyperAudio Gateway

We developed an alternative player whose user interface device is a telephone. The player consists of a telephone on Public Switched Telephone Network (PSTN) and a HyperAudio gateway between Internet and PSTN.

To listen a HyperAudio content, a user calls a HyperAudio gateway through PSTN and requests a content by pressing buttons, at first. The gateway receives the request and gets the ILSL file from the ILSL server. Then the gateway analyzes it to determine which audio data to be played, sends requests to the audio server, receives the audio data, and plays them. To notify hot spot portions for hyperlinks, special tones are played at the start and the end of hot spots.

4.4. HyperAudio Contents

For contents, news and guide book contents are being tested. They can be listened to either sequentially from the start to the end, or not sequentially for getting specific topics directly.

So far, we have not concluded what kind of contents and what kind of structures are most suitable for the listening. We should try to several kinds of contents and structures in future.

5. SUMMARY

In this paper, we proposed the concept of the interactive listening of non-sequential audio contents (HyperAudio). We have implemented a prototype system consisting of a HyperAudio server and HyperSpeech players connected via Internet.

6. REFERENCES

1. Berners-Lee, T. and Connolly, D. Hypertext Markup Language – 2.0, Internet-draft, IETF, 1995.
2. Raggett, D. HTML 3.2 Reference Specification, W3C Recommendation 14-Jan-1997, W3C, 1997.
3. Raggett, D., Le Hors, A., and Jacobs, I. (eds.) HTML 4.0 Specification, W3C Recommendation 18-Dec-1997, W3C, 1997.
4. Raman, T. V. Auditory user interfaces, Kluwer Academic Publishers, 1997.
5. IBM. Home Page Reader Manual, Version 1.0, 1997.
6. Bos, B., Lie, H. W., Lilley, C., and Jacobs, J. (eds.) "Aural style sheets", CSS2 Specification, W3C Working Draft 04-Nov-1997, W3C, 1997.
7. Tim Bray, T., Paoli, J., and Sperberg-McQueen, C. M. Extensible Markup Language (XML) 1.0, W3C Recommendation, 10-February-1998, W3C, 1998.
8. Bugaj, S. et al. Synchronized Multimedia Integration Language (SMIL) 1.0 Specification, W3C Recommendation 15-June-1998, W3C, 1998.
9. Fielding, R., Irvine, UC, Gettys, J., Mogul, J. Frystyk, H., andBerners-Lee, T. Hypertext Transfer Protocol – HTTP/1.1, RFC 2068, IETF,