Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan, July 2-5, 2002

LISTENING POST: GIVING VOICE TO ONLINE COMMUNICATION

Mark Hansen

Bell Laboratories, Lucent Technologies
600 Mountain Avenue, 2C283
Murray Hill, NJ 07974
cocteau@bell-labs.com

Ben Rubin

EAR Studio
111 Bowery, 3rd Floor
New York, NY 10002
benrubin@earstudio.com

ABSTRACT

Listening Post is a multimedia art installation designed to convey the magnitude and diversity of online communication. This unique space provides a meaningful rendering of a massive data stream consisting of thousands of simultaneous conversations. In this article, we explore specifically how the audio component of the installation provides a structure that enables visitors to make sensible inferences from this complex, dynamic data. *Listening Post* makes use of a multi-layer audio display consisting of mechanical noises (relay clicks), sampled sounds, and synthesized voices. We illustrate how these components, together with a very simple visual display, combine and interact to give visitors a sense of the topics being discussed in thousands of chat rooms, all in real time. Finally, we discuss some of the systems and software infrastructure necessary to create the complex audio display.

1. INTRODUCTION

The advent of online communication has created a vast landscape of new spaces for public discourse: chat rooms, bulletin boards, and scores of other public on-line forums. While these spaces are public and social in their essence, the experience of "being in" such a space is silent and solitary. A participant in a chat room has limited sensory access to the collective "buzz" of that room or of others nearby – the murmur of human contact that we hear naturally in a park, a plaza or a coffee shop is absent from the online experience. The goal of our project is to collect this buzz and render it at a human scale. We use sound, text, motion and space to create sensual encounters, abstracting the communication spaces away from their familiar on-screen presence.

Taken together, public online communications represent an enormous outpouring of real-time data, and this data is filled with complex structure. Topics emerge in response to current events and daily activities in cycles that vary hour-to-hour, day-to-day, and season-to-season. The emergence of these topics transcends the boundaries of the online landscape: a local knitting circle in Australia and a political discussion group on Yahoo may both react to news of a political scandal or a world cup victory. Our goal is to distill the content and the structure of this collective communication and to present it in ways that are accessible and compelling.

Listening Post is an art installation created for public presentation at the Brooklyn Academy of Music from December 6-20, 2001. During that period, roughly 1,800 people saw the piece. Background material about the project appears in [1]. Audio recordings from the installation can be found in [2]. *Listening Post* emerged from a collaboration between the authors as part of a larger project to explore the intersections between statistics research and sound art. See [3] for a details.

2. RESEARCH GOALS

Our starting point for *Listening Post* is an enormous stream of public Internet chat rooms, bulletin boards and online forums (which we will collectively refer to simply as forums). These sources distinguish themselves from other content on the Web in that they involve ongoing exchanges between two or more people. Variants of the software monitors described in [1] are used to cull material from up to 5,000 forums in real time, focusing attention on the most active sources. Fundamentally, *Listening Post* is an attempt to understand the patterns that emerge from thousands of simultaneous conversations and the dynamics that govern their shifting topics. As a rendering device, the installation space is designed to convey several aspects of the data stream collected by our software monitors. These concepts shaped both the physical construction of *Listening Post*, as well as the underlying audio, visual and technical vocabularies that govern how data are presented.

- **Content** Perhaps our most important goal is to give visitors a sense of *what people were talking about.* The conversations taking place in each online forum were largely independent, and yet existed in the same social context (current events, pop culture). By looking at thousands of rooms, patterns emerge that relate to how people communicate.

- **Scale** At any point during the day, our content monitors capture the contributions of tens of thousands of people. It is not possible to render all these data in real time, and yet *Listening Post* has to convey the scale of the system being observed.

Immediacy The forums under study are inherently dynamic, some becoming active only at night, while the topics in others change over the course of the day. Visitors to *Listening Post* need to have a sense of the dynamic nature of these forums, and that if they return in a week's time, the subjects will be different.

3. PREVIOUS EXPERIMENTS

In [3], we considered a technique for sonifying the conversations in a single chat room. While our previous experiments with network data were based entirely on a tonal representation of data streams, we found that an audio display for chat had to incorporate a spoken component to adequately convey content. For the sonification of a single chat room, we began mixing synthesized voices from a text-to-speech (TTS) engine into our sound design. In [3], these

voices presented the popular topics in the room, and periodically read representative posts verbatim from the chat stream. Algorithmically generated piano accompaniment differentiated between on-topic and off-topic conversations.

In a public performance at The Kitchen [4] in New York City (April, 2001), we experimented with presenting up to 50 forums at once. Replicating our recipe for a single chat room, we attempted to have up to four simultaneous voices reading content from different forums, each voice coming from a separate speaker placed in the four corners of the performance space (separated by 30 feet). As expected, spatial separation seemed to improve the listeners' ability to understand what was being said. We made further gains by assigning different pitches to voices from different speakers. Subjectively, we also found that monotone or chanting voices were easier to separate, and provided a more cohesive mix. Although these elements appeared effective, we observed, not surprisingly, that when we added a projected visual display of the four text streams, the audience was much better able to attend and comprehend the spoken text. When considering scaling the Kitchen experiment from 50 rooms to 5,000, this final observation led us to design a visual element for the *Listening Post* display system presented here.

4. AUDIO AND VISUAL COMPONENTS

The visual centerpiece of the of Listening Post is a 7' by 10' array of 110 small text displays, each measuring 2'' by 6'', and able to hold four lines of 20 characters. A sample of a single display is given in Figure 1. The displays are arranged in 10 columns, each with 11 rows, and the individual displays are set 6'' apart vertically and horizontally (See Figure 2. Here the text displays are in "big character" mode where the original 4x20 characters are used like pixels to show up to four big characters; in this mode content scrolls from right to left). While the array is simple in conception, its size makes it difficult to navigate visually. Even with only half of the displays active, visitors find it difficult to quickly summarize or make broad inferences about the content being displayed on the array.

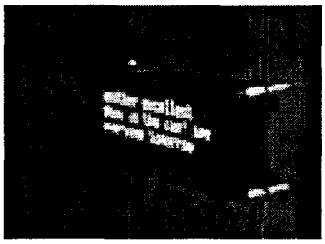

Figure 1: A single display unit.

Sound, then, is essential if we expect visitors to extract meaning from the installation. The physical sound design has several main features:

- Multiple sound sources: Ten speakers (plus a subwoofer) carriy the spoken audio, tones, and other sampled or synthesized audio.

- Mechanical point sources: Mounted to the back of each display unit is a relay which can be actuated under software control. By varying the duration of the relay actuation pulse, we can control the loudness of a mechanical "click" from each display, in effect creating a 10x11 grid of 110 controllable point-sources for these clicking sounds. Each display can then make a click that was loud enough to be heard over the other sounds in the room, drawing a visitor's attention to that spot on the array.

In the end, the visual component of *Listening Post* acts as a kind of ventriloquist's dummy, which is animated by our sonic design. Before we elaborate on this point further, we briefly describe the acoustics of the installation space.

Figure 2. The display grid.

5. PHYSICAL LAYOUT AND ACOUSTIC DESIGN

The aural impact of our display depends on having a controlled acoustic space. The room itself measures 17' by 32' and the layout is given in Figure 3. The floor is carpeted, and the walls of the room are covered with 28 panels of perforated aluminum. Behind each panel is a 2''-thick fiberglass acoustical panel. A baffled entrance reduces the transmission of sound to and from the corridor outside the room. These acoustical treatments yield a relatively quiet and non-reverberant room, enhancing intelligibility. 8 of our 10 speakers are mounted behind the aluminum panels, and the remaining 4 (as well as the subwoofer) hang high up and out of view.

6. PRESENTATION STRUCTURE: "SCENES"

A visitor to this installation experiences four different scenes or movements in a cycle that last about 10 minutes. Each scene is constructed around a different display algorithm; the mapping and dynamics of sound, visual elements and data are unique to each scene. In this section, we describe one such scene in detail, and will present the others in the full version of the paper.

The scene we chose to describe is designed specifically to highlight content. Because of this, it is the most complex of the four that comprise *Listening Post*. This scene is best characterized in terms of an agent that manipulates text and effects changes in the audio/visual display. An agent associates itself with one of the locations in the array. At the beginning of the scene, this choice is random. When the agent chooses a location, text will scroll rapidly on the chosen display.

During this scrolling, the agent is examining the stream of posts to the forums we are monitoring in an attempt to find one that "matches" the text displayed in the screens around it (or, if there is no text around a given display, the agent will attempt to find the next entry that matches the text of an post chosen at random from the stream). The sense of a match is best described in terms of classical information-retrieval metrics popular in many Internet search engines [5] and will be documented elsewhere.

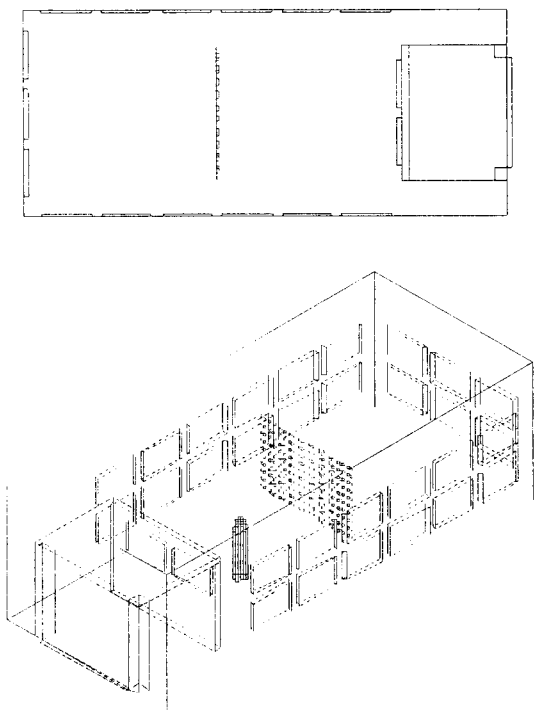

Figure 3: Plan and isometric views of *Listening Post* installation room, overall dimensions 17' x 32'. The display grid hangs in the center of the room, and acoustic panels are mounted on all wall surfaces.

While the agent searches in this way, the screen refreshes 10 times per second, each time displaying the next post in the stream (the current target of the matching test). Each new message being scanned is announced by a low-volume relay click. With the 10 Hz refresh rate, the series of clicks becomes a whir or a fluttering sound. In the world of this scene, that fluttering sound is tightly linked to this searching process. If a match is found within 20 seconds the scrolling/fluttering stops, a loud click is heard from this display unit, and a single message is held on the display. Simultaneously, the agent triggers a pitched sample to be played in the room (in this case a sampled bell or gong) and a monotone voice with the same pitch intones the text on the display.

The text will then disappear from the screen after a minute. Before this happens, the agent jumps to an empty neighboring screen (if one is available) and starts the searching all over again (this time looking for a match among its new neighbors). If no match is made after 20 seconds, the agent simply jumps at random to another part of the array and begins a new search. Over the four-minute duration of this scene, more and more agents appear. By the time a dozen or more agents are working simultaneously on the array, the irregular patter of clicks, tones and voices takes on an arrhythmic musical pattern. The layers of pitched voices take on the quality of a chant or litany as they blend with each other and with the reverberating tones. By the end of the scene, as many as forty voices can be heard and about 2/3 of the 110 displays show messages.

This agents' activity provides a kind of self-organizing structure to the content displayed on the array. Topics that are common in the stream at that time start to consume large segments of the array. During the presentation at BAM (December, 2001), topics like "terrorism" or "Afghanistan" could take up a quarter of the display. Persistent chat room pleas for "single men" or "single women" with certain physical characteristics appear and cluster regularly throughout the display. By using solely IR-related match statistics, text from different sources can be strangely juxtaposed; keying off of the work "hot" one might find posts related to global warming or singles chats.

While the agents provide a visual organization to topics on the array, it is sometimes difficult to assess content from reading the individual panels. Some visitors step back from the array, preferring to listen to the soundscape and watch (without reading) the pattern of light made by the agents as they scroll through the text.

As the agents move around the screen, their activity is made meaningful through our audio design. Using Gaver's terminology [6], the audio engages listeners in terms of both *everyday listening* (the clicking of the relays drawing attention to the location of the sound sources) as well as *musical listening* (the overall soundscape forming organized patterns of pitch, harmony and rhythm). The pitched voice and tone reinforce the material on the screen, allowing the visitor to extract the content on the screen with a mix of reading and listening.

The sequence of tones selected also follows a kind of self-organizing principle. The pattern that emerges is musical, and the visitor can begin to predict likely "next pitches". This prediction is not completely accurate, however, because we are not simply cycling through a fixed sequence of pitches. The prevalence of topics fuels the rate of matches which in turn drives the rhythm and mix of pitches in the room.

Subjectively, it seemed as though it was possible to selectively attend to a given utterance even when there may have been as many as forty simultaneous voices in the room. This degree of stream segregation was surprising to us, and we suspect it is due to a combination of several factors:

- The pitching of the voices and the musically-arranged sequence in which pitches were chosen;
- The creation of compound events composed of three types of sounds together (click, tone, voice);
- The use of multiple sources of sound in space.

Through its mixture of audio, visual and spoken content, each of the scenes designed for *Listening Post* required a kind of "polyattentiveness" [7] from the visitors. We sought to achieve our goals of conveying content, scale and immediacy by a careful orchestration of audio, spatial relationships, acoustic conditions, and supporting visual cues.

7. SYSTEMS AND SOFTWARE

The audio and visual display for *Listening Post* is driven by a network of 3 PCs and a Mac G4. One PC (Windows NT) is responsible for the material displayed on the array of text screens. This PC runs a number of drivers that also control the kind of sound made by each display (click presence/loudness). One of the PCs (Windows 2000) runs a commercial TTS engine [8] that produced as many as forty voices in the room at one time. The Mac runs Max/MSP [9] and is responsible for all the other sounds generated in the room. The final PC, running Linux, acts as a kind of coordination engine that orchestrates communication between the various audio and display components.

Because of its power as a compositional tool, Max is responsible for determining values for aesthetic parameters governing the audio display. For inter-process communication, we make extensive use of the Open Sound Control (OSC) [10] protocol originally developed for Max. We wrote a general purpose OSC client in Perl [11] so that Max can communicate with the other pieces of the system. We created a sequence of OSC devices that specify scene type and parameter values. Messages sent to Max included start/stop indicators for the scenes, tickers to record specific events within each scene, and activity monitors that kept Max informed of the activity on the display.

As an example, consider the "content" scene explained in Section 4. When an agent identifies a sample to display, several events are triggered simultaneously: 1) the relay on the display makes a loud clicking noise; 2) Max generates a pitched tone, and 3) the TTS engine reads the content displayed on the screen in a monotone voice pitched to match that of the introductory tone. Here is the sequence of events that take place to create this. To start the scene, the controller on the Linux PC sends an OSC message to a port on the NT computer corresponding to this scene. The message specifies how long the scene should run for. When the scene starts, the program on the NT computer sends Max an OSC message indicating that the scene has begun. It also starts a single agent scrolling on the display, and gradually introduces more as the scene progresses.

When Max receives notification that this scene has begun, it sends the TTS engine an OSC message specifying the pitch and volume that the next voice should speak at. These messages are of the form `/lp/content/pitchvol p v` where p and v are integers. (In terms of the OSC protocol, the first string is a symbolic address that we structure to represent the project, `lp`, the scene, `content`, and the parameter names, p and v.) When one of the agents finds a match in the data stream, it sends the message to the display along with the specification that a loud click be issued. It also sends signals to Max and the TTS engine. The latter message consists of the text the TTS engine is to speak (at a pitch and volume previously specified by Max). The notice to Max is of the form `/lp/content/pulse`. Periodically, Max will also receive messages that record the "activity" on the display; that is the number of text units that currently hold content. Max uses this to adjust the volume of the voices in the room. The OSC message is now of the form `/lp/content/activity a`, where a is an integer from 0 to 110. When Max receives notice that a match happens, it plays a sample with the pitch sent previously to the TTS engine. It also sends an OSC message to the TTS engine, giving it the volume and pitch of the next voice.

As can be seen from this example, each of the computers involved in creating a *Listening Post* scene speaks more or less directly to each of the other computers. OSC is the substance of this communication.

8. CONCLUSIONS

In this extended abstract, we have documented one of the four scenes of *Listening Post*, our attempt to convey the scale and content of thousands of conversations in real time. We have explained the interplay between localized, mechanical sounds, musically arranged tones, and pitched voices in designing a display that reflects the dynamics of online communication. The combination seems to provide us with an ability to separate one out of a stream of many voices. This has the effect of conveying both scale (the impression of multiple voices talking at once) and content (by isolating a single voice, we can hear one person's contribution to the stream of thousands). In future extensions of *Listening Post*, we are planning to monitor even more sources and render the data on an array consisting of even more displays. As part of our preparation for these extensions, we plan to conduct experiments to understand how the different audio components aid our ability to separate voices and whether other enhancements might allow us to better convey content. In future incarnations of *Listening Post*, we intend to add even more scenes, highlighting different aspects of the data stream. The interested reader can learn more about this project at the Web site [12].

9. REFERENCES

[1] Mirapaul, M. Making an opera from Cyberspace's Tower of Babel, *New York Times*, Monday, December 10, 2001. www.nytimes.com/2001/12/10/arts/music/10ARTS.html

[2] Studio360, a production of WNYC. January 24, 2002. www.wnyc.org/new/Studio360/show012602.html

[3] M. H. Hansen and B. Rubin (2001) Experiencing information systems through sound, *Proceedings of ICAD 2001*, Hiipakka, J., Zacharov, N. and Takala, T., Eds., pp 10-15.

[4] The Kitchen. www.thekitchen.org

[5] Frakes, W. and Baeza-Yates, R. (1992) *Information Retrieval: Data Structures and Algorithms*, Prentice Hall.

[6] Gaver, W. (1993) How do we hear in the world? Explorations in ecological acoustics, *Ecological Psychology*, 5, pp 285-313.

[7] Sayre, H. (1989) *The object of performance: The American avant-garde since 1970*, University of Chicago Press, Chicago/London.

[8] The Articulator TTS engine, Lucent Speech Solutions. www.lucent.com/speech

[9] Cycling 74. Max/MSP, www.cycling74.com.

[10] M. Wright. Open Sound Control. cnmat.Berkeley.edu.

[11] Wall, L. Christiansen, T., and Orwant, J. (2000) *Programming Perl*, O'Reilly & Associates, Sebastopol, CA.

[12] www.earstudio.com/projects/listeningPost.html