Visualizing Remote Voice Conversations

Pooja Mathur
University of Illinois at Urbana-Champaign, Department of Computer Science
Urbana, IL 61801 USA
pmathur2@illinois.edu

Karrie Karahalios
University of Illinois at Urbana-Champaign, Department of Computer Science
Urbana, IL 61801 USA
kkarahal@illinois.edu

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.
ACM 978-1-60558-247-4/09/04.

Abstract
Online voice conversations are becoming ever more popular. People have been logging online text conversations, but what about voice conversations? Walter Ong simply states, "written words are residue. Oral tradition has no such residue or deposit" [6]. However, we do not just want to archive conversations, we want to enable users to have some meaning in these "logs". We introduce a project that takes a remote conversation and visualizes it. It does so in a way that takes volume, pitch and content into account. With this information, the visualizations display the data in a meaningful way. Users can use these images in the future to review past conversations whether it is for nostalgia's sake or to recall some piece of information. In this paper, we describe the early design and iteration of system for archiving and creating artifacts from remote audio conversations.

Keywords
Conversation, social visualization, Skype, VoIP, archival, artifacts, remote audio, content, volume, pitch

ACM Classification Keywords
H4.3 Information Systems Applications: Communications Applications – Computer Conferencing, teleconferencing, and videoconferencing. H5.m. Information Interfaces and Presentation (e.g., HCI): Miscellaneous.

Introduction

Sound is ephemeral. After speaking, the sound waves dissipate and the speaker is left with nothing but the mere memory of what was said. How one could be expected to remember everything that was said is unknown. Whether we should archive everything for posterity is also an issue for debate. For some time, people were limited by storage space even if they were capable of logging speech. Storage space is less of an issue now, but who has the time to a review hours of logged speech to recall some piece of information? There is need for a way to log conversations without making the user spend hours reviewing the logs.

Without sound, arts of speech, like sarcasm can be lost. Numerous people have tried to use sarcasm in a text conversation and as a result, offended the person they were speaking to. Furthermore, auditory signals are not the same in text conversations. Can the essence of a laugh really be captured in the three letters, "lol"? The audio of a conversation can depict sarcasm, question, and subtle cues in a manner that a transcript of audio cannot.

An interesting question was brought up by Hollan and Stornetta [3]; in their work they ask "what would happen if we were to develop communication tools with a higher information richness than face-to-face"? What if we could also use audio conversations as artifacts of our lives? People bring back souvenirs from all around the world to remind themselves of the places they have been. This reflective property [5] can also be applied to conversations. We could take a picture of a conversation and use that image to remind oneself about the topic of the conversation, the people in the conversation and more.

To make a picture out of the conversation we need to visualize the information. This could include the volume, pitch, and content of the conversation, among other features. In this paper, we describe a new project that attempts to take various aspects of conversation (volume, pitch, and content) and display it in a meaningful way.

Related Works

One of the earliest projects visualizing aural conversation started out by representing the virtual space in which the conversation was held on the screen. In the Somewire [8] project, one of the four interfaces focused on the social aspect of audio communication. Called Vizwire, this was one of the first projects that allowed its users to define the social space though visual cues. Users were represented by icons and could move their icon around in the space. Then, in Talking in Circles [7], users were abstracted as circles in this virtual space. Users could move their circle around and participate in different conversations. A user could only hear the speech from other users whose circles were within a certain radius of one's own.

In a later project, Karahalios and Donath took the idea of visualizing audio space further. The Telemurals [4] project connected two remote spaces. When a user stood in a certain area, a faint outline of the user would be projected at the other Telemural site. If the user spoke, the faint outline would start to color in, and more details would be shown. Using the amount of coloring and detail as a cue, users could note user participation in the conversation between the spaces.

Visiphone [2] also connected two remote spaces. Each space was given a color, and the audio signal is viewed

over an interval. A circle appears on the display in proportion to the audio signal over that interval. From this, users are able to gather presence, volume, and a short history. The Conversation Clock [1] took a different route and visualized collocated group conversation about a table. Each user is given a color; when that user speaks a tick mark of that color appears on the display area. From this users can notice many things. Among which, users can get an idea who is actively involved in the conversation and who is not.

Apart from audio, Themail [11] is a project that focuses on content extraction. In this project, information from email is taken to give users an idea of what they discuss with others and how the themes prevalent between some contacts are different that those between others. Themail takes this information and displays the information to show how topics change over time.

Current Work

Our current work focuses on dyadic conversations through the application Skype [9]. While our system, titled VoiceSpace, is still in development, our intent is to create a plug-in for the Skype program. This plug-in will grab the content of the conversation using the speech APIs available through Microsoft Windows Vista.

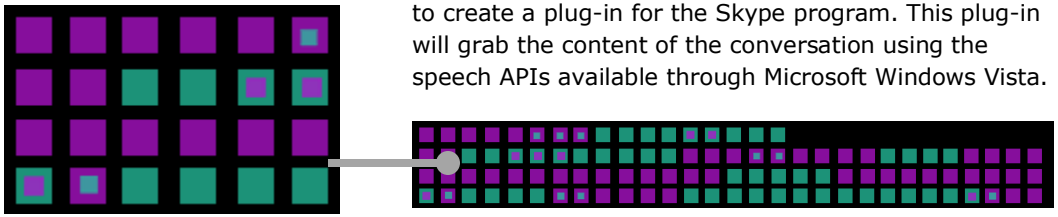

Figure 1b. This is a close-up of history view. The sizes of the squares are about the size that will appear on the screen.

Figure 1a. This image is a mock up of the History view. The conversation has been going for about 3.5 minutes.

This project contains three different visualizations. One of the visualizations focuses on the history of the conversation. The next visualization focuses on the volume and pitch of the conversation. The last visualization focuses on the content of the conversation.

History View

The History View visualization is considered the baseline of the all the visualizations. Both users in the conversation are given a color. When a user speaks, a square of that user's color appears on the screen. If users speak over one another, the quieter of the two gets a smaller square that appears inside the larger square, as in Figure 1b.

The squares appear along the bottom horizontal edge of the window, from left to right. Each horizontal line of squares represents one minute of the conversation. The first minute of the conversation starts at the bottom left corner. Then each minute after appears a few pixels above the last. See Figure 1a for a mock-up with a few minutes of conversation history built up.

Through this visualization, we imagine that the users can gain a number of insights. First, users will be able to get an idea of who controls the conversation. By looking at which color appears more often, users can tell who was talking more often.

The visualization can also give users insights into backchannels. If one participant is speaking the other participant may say things like "I see", "Hmmm", or "Really?" to show the speaker that they are listening. The speaker can then not only hear that the other participant is listening, but one can see it as well. The smaller inner squares might be representative of backchannels.

Figure 2b. This is a close-up of pitch and volume view. The sizes of the circles are about the size that will appear on the screen.

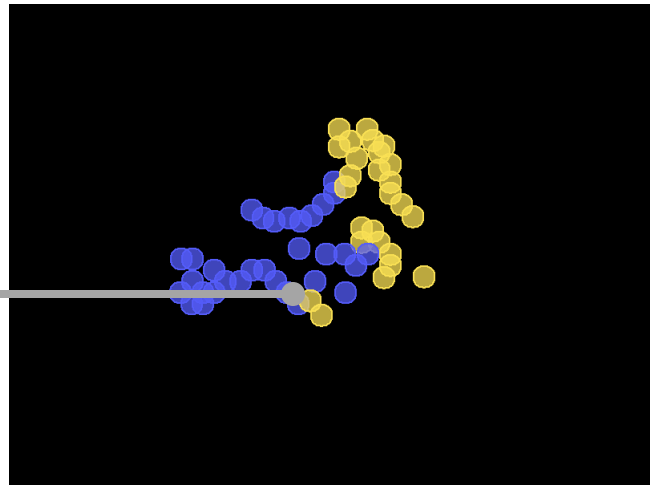

Figure 2a. This image shows the mock-up of the pitch and volume view. If the x-axis is considered to be pitch and the y-axis is considered to be volume, than we can see that one user talks with a lower pitch and the other participant speaks with a high pitch, on average.

Pitch and Volume View

The next visualization is designed as a space for users to explore, rather than to simply report back information about the conversation. In this view, users can explore the relation of their volume to their pitch.

As in the first visualization, each user gets assigned a color. When a participant speaks, a circle appears on the screen. The circle's location is dependent on the user's volume and pitch. The x-axis graphs the user's pitch. The left end of the screen represents a low pitch and the right end represents a high pitch. The y-axis graphs the user's volume. The bottom of the screen represents a low volume and the top of the screen represents a high volume. In the implementation of VoiceSpace, we plan to make these characteristics interchangeable with the axes to allow the user to view the space in the way that is most intuitive to oneself.

In this visualization, users can potentially reveal patterns in their speech. For example, if one participant laughs often, the same pattern of circles often appears on the screen. Users can discover which participant speaks louder or softer on average. In the mock-up shown in Figure 2a we can see that one user was speaking with a lower pitch and on average slightly softer. The other participant was speaking with a higher pitch and was slightly louder on average.

Content View

This last visualization gives users the opportunity to view the content of their audio conversation in a meaningful way. We did not want to simply transcribe the conversation, wanted to pull out the salient information. With this in mind, the initial idea for this visualization was to appear like a tag cloud. However we wanted the data to appear chronologically as it was used in the conversation.

Figure 3a shows a mock-up of the visualization. Words from the conversation start appearing in a circle around the center of the screen. Like the other visualizations, both users have a different color used to represent their information.

Each circle of words represents a minute of the conversation. After a minute is finished, the words that are currently on the screen move in an outwards direction from the center of the screen to make room for content of the current minute. Then the next minute starts and new content begins to appear.

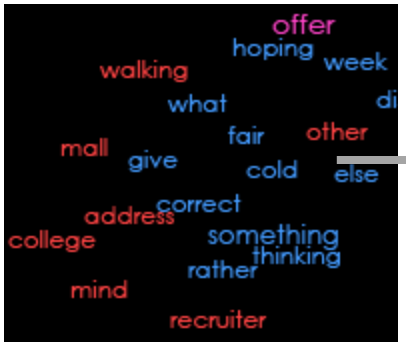

Figure 3b. This is a close-up of content view. Note that the word "offer" is a mix of the two main colors. The font size is also slightly larger than many of the other words on the screen.

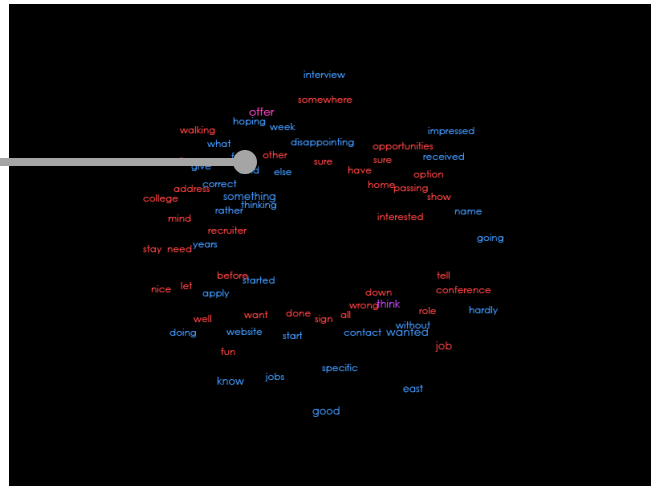

Figure 3a: In this mock-up, we get an idea of what the visualization may look like after about 5 minutes of conversation. See close-up for more detail.

The idea of the tag cloud comes in when words are repeated. For example, if the participants in the conversation were talking about finding jobs, they may use the word "offer". Every time the word "offer" is repeated, instead of rewriting the word on the screen, the first location of the word is found and the font size and brightness level are increased. If the word is said by the same speaker that spoke the word initially, then the color does not change. However, if both have spoken the word, then the color becomes a mix of both of the main colors. Furthermore, every time a word is said, it moves closer to the center of the screen.

Users can now not only get an idea of turn-taking and dominance, but they can also get an idea of what themes are present in their conversation. Words that appear large and closer to the center are words that were used often in the conversation.

Artifacts and Archival

After completing a conversation, users are left with an abstracted log of their conversation. With an image of the content view, users can come back and look at the image to recall the content of the conversation without having to read an entire transcription or listen to an audio file. Not only does this type of logging save space, but it also saves time. A user can keep these images as an archive of their audio conversations just like those of text conversations. Users have an option of which of the three visualizations to look at.

Any of the views could be used as an artifact of a conversation and one's relationships [10]. A history view that exhibits the participants were often talking over one another could remind the user of an argument. The pitch and volume view can display a conversation of excitement if the users' circles often appear near the right side of the screen. The content view can easily be used as an object to recall topics of past events. Each of these visualizations has the ability to stir emotion and memories.

Future Work

The next step in this work is to complete the implementation. As stated earlier, we are using the Skype APIs and the speech APIs that come with Microsoft Windows Vista to get the audio over the network and analyze it. After getting the backend set up, we will work on creating the visualizations. The mock-ups presented in this paper are early designs. We expect the visualizations to go through several iterations before their final implementation and design.

We will then perform user studies to see how typical Skype users feel about them. We want to see if the visualizations can be used for archival and how they are used as artifacts.

With remote audio conversations becoming increasingly popular, the need, or desire, for quick and easy archival and information retrieval may also follow in suite. There is also the growth in popularity of social media. People may start posting screenshots of their conversation visualizations on their Facebook page to share with others. It will be exciting to see what new uses and meanings for the visualizations our participants will come up with.

Acknowledgements

We would like to thank everyone in Social Spaces at Illinois, NSF, and Microsoft for their input and support of this work. We would also like to thank members of the Skype Developers Forum for helping us understand the Skype APIs further.

References

[1] A. Bergstrom and K. Karahalios. Conversation Clock: Visualizing audio patterns in co-located groups. *Proc. HICSS 2007*, IEEE Computer Society (2007), 78.

[2] J. Donath, K. Karahalios and F. Viegas. Visiphone. ICAD 2000.

[3] J. Hollan and S. Stornetta. Beyond Being There. *Proc. CHI 1992*, ACM Press (1992), 119-125.

[4] K. Karahalios and J. Donath. Telemurals: Linking Remote Spaces with Social Catalysts. *Proc. CHI 2004*, ACM Press (2004), 615-622.

[5] D. A. Norman. *Emotional Design: Why we love (or hate) everyday things*. Basic Books, New York, 2005.

[6] W. Ong. *Orality vs Literacy*. Routledge, New York, 2002.

[7] R. Rodenstein and J. S. Donath. Talking in Circles: Designing a Spatially-Grounded Audio-Conferencing Environment. *Proc. CHI 2000*, ACM Press (2000), 349.

[8] A. Singer, D. Hindus, L. Stifelman and S. White. Tangible Progress: Less is More in Somewire Audio Spaces. *Proc. CHI 1999*, ACM Press (1999), 104-111, 625.

[9] Skype. http://www.skype.com/

[10] F. B. Viegas, d. boyd, D. H. Nyugen, J. Potter, and J. Donath. Digital artifacts for remembering and storytelling: PostHistory and social network fragments. *Proc. HICSS 2004*. IEEE Computer Society (2004).

[11] F. B. Viegas, S. Golder and J. Donath. Visualizing Email Content: Portraying Relationships from Conversational Histories. *Proc. CHI 2006*. ACM Press (2006), 979-988.