Jonathan Bruck
3.19.1998

Proximal Control Through 3D Audio

This paper documents my search to find ways to create more intimate and personal experiences for voice interfaces. This basic assumption is that proxemics plays are large part in social communication and these cues were non-existent in current voice interface. I then discuss proxemics in detail and the importance it plays in our physical world and how new technology in 3D audio is able to reproduce these audio cues. Finally, I merge proxemics and 3D audio to discuss the implications for controlling virtual spaces through the use of 3D audio in technology.

What Are Proxemics?

Voice interfaces have amazing potential and are sure to greatly enhance people's efficiency and interactions with in the future. However, a necessary step in the future for designing a successful voice interface will be to account for the role proxemics play in social interactions. Proxemics can be summarized as the study of spatial distances between individuals in different cultures and situations. Proxemics is important to voice interface because it is a main component people use for successful human-human interactions and thus is necessary for equally successful human-computer interactions. The auditory sense plays an important role in proxemics because sounds are responsible for giving people cues about physical location and distances of people and objects in their surrounding environment. Most of the major research in proxemics was done by Edward Hall in the 1960's, in which he developed the four zones of proximity. He defines these zone as the intimate (1-18 inches), personal (18in - 4ft.), social (4-12ft.), and the public (12+ ft.) (Hall, 1966). The one drawback to proxemics is that every culture tends to have different rules for their four zones of proximity. However, because voice interfaces are also cultural dependent, I will be focusing on the effects of proxemics within America.

The intimate zone is usually reserved for only the closest people in our lives, lovers, friends, and family. At the nearest distance between 1 to 6 inches, sound is used as reinforcement that the person is close, and the primary form of communication in this area is conveyed by touch. However, beyond 6 inches the voice becomes important again in communicating distances and location. When a person is speaking to someone in their intimate zone the voice is usually very close to their head and is in a very low voice or even a whisper (Hall, 1966), the obvious exception being when two people are fighting. The second zone or personal distance is "the term used to designate the distance consistently separating the members of non-contact species. It might be thought of as a small protective sphere or bubble that an organism maintains between itself and others" (Hall, 119). The extent of this zone is approximately four feet or the distance between two people when one person's hands are stretched out touching the other. Again, this zone is mainly reserved for people who are close to us and the voice level is moderate.

Our third zone is the social distance 4-12ft. In this zone the majority of business is conducted. Closer for more informal business and coworkers, while farther out on the spectrum for formal business. Also, this is a common distance for casual social gathering. At the far end, beyond 7 feet the voice is noticeably louder and "can have the effect of reducing social distances to personal distance. (Hall, 123). Finally, the fourth zone deals with our public space of 12 feet and out. Since this zone has no real end, the most important area is between 12 and 20 feet because the human ear loses efficiency beyond twenty feet. In this range the voice is loud but still not at full volume. Additionally, "linguists have observed that a careful choice of words and phrasing of sentences as well as grammatical or syntactic shifts occur at this distance" (Hall, 123).

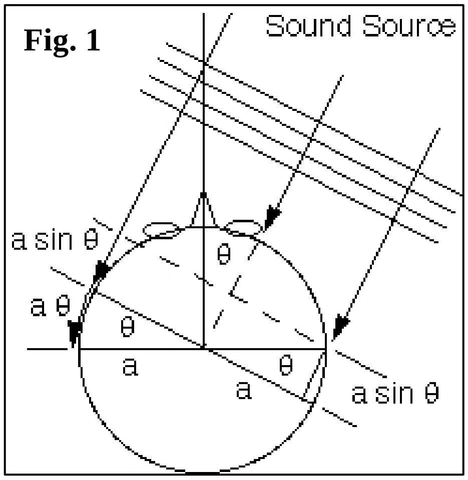
Proxemics are important because everybody follows the same rules within a particular culture. Thus, the implications which are discussed later are numerous, because each zone is applicable to specific interactions that voice interface will be dealing with. Thus, if it is possible to control distance and location cues artificially through stereo speakers, it should be possible to control proximal rules that will enhance specific interactions in social situations.

3D Sound

Since proxemics plays an integral part in human-human interaction the same responses should occur in human-computer interaction. This assumption led me to my next step which was researching if and how technology could control proximal cues through audio. The search was surprisingly rewarding as I stumbled upon a half a dozen companies that are dedicated to 3D audio. Out of all the companies, one stood out above all the rest in their application of developing 3D audio. The company is called SRS Labs (www.srslabs.com) and while it is not necessary to go into detail about the company and its products in this paper, their research and development of 3D audio should be monitored.

3D sound differs dramatically from stereo sound which at best provides a left-right panning effect and is generally restricted to a one or two dimensional plane. On the other hand, 3D sound seems to originate outside of the listeners head in three dimensional space. It can place sound above or below, to the left or right, and in front or behind a person (Intel Corp., 1997). On top of this, 3D audio is able to accomplish this with the use of only two speakers. 3D audio is based on the principals of how the human ear hears. These principles have been broken into three distinct areas which are responsible for generating 3D audio in our environment. The first major cue for 3D sound is called the Interaural Time Delay (ITD) and is largely used to determine left vs. right localization. The ITD is the difference in time it takes a sound coming from a certain direction to hit your ears (Figure 1). For example, a car's horn honking to your left would

Fig. 1



hit your left ear a brief second before it can get around your head and into your right ear. While, this is a strong cue, we are also dependent on what is call the Interaural Intensity Difference. This works in conjunction with the ITD and is the difference in sound intensity between the two ears. Thus, the same car horn would be louder in the right ear than the left because the head dampens the sound. This dampening affect is called the head-shadow effect (Duda, 1997). With these two cues we can very accurately locate sounds coming from the left and right. The third major cue our ears use for 3D sound is called Head Related Transfer Functions (HRTF). The HRTF captures all of the physical cues to source localization by accounting for the spectral shifts caused by a variety of physical features. These are "the shape of the pinna (outer ear), the absorption of sound by our head, hair, and body, sound reflections off our shoulders, etc." (Pulido, 1997). The HRTF is largely believed to be responsible for front and back localization. Important to note the HRTF is far more complicated than reported here, but as this paper is dedicated to 3D sound use in voice interfaces, further explanation is not necessary.

SRS's technology essentially mimics these diffractive effects by extracting information from a recording that originally came from the sides and rear in a physical environment. SRS then uses HRTF based corrections to cause the ear to perceive these sounds in their original spatial relationship. Thus, SRS is essentially able to fool our ears with two speakers by reproducing the natural frequencies the sound would have made in a physical environment.

Implications

Based on the research and first hand experience with 3D sound, I feel that it is highly feasible and realistic to reproduce natural acoustic environments through two speakers. Thus, it should be possible to control the four zones of proximity in virtual environments though the use of 3D sound. Being able to control the four zones of proximity in voice interfaces and in technology begs for further research of the full implications 3D audio has to offer in creating more real, personal, and intimate experiences with technology. Below I outline a few key areas where 3D audio can immediately be applied to control social proximity in voice interfaces and computers.

Privacy issues with Speech to Text and Text to Speech systems.

One major area developing in voice interfaces is Speech to Text (STT) and Text to Speech (TTS) applications. These applications will allow a person to talk directly to the computer and visa-versa. For example, a person will not be required to type an email message anymore and instead can just speak the message while the computer is converting the speech into text. Many companies feel that this will be a killer application for computers. However, I feel that the current model will not hold up because issues of privacy and eavesdropping will be greater than anticipated which would dissuade people from using such systems. Only in a closed environment would a person feel comfortable talking out loud to their computer, but many work environments today are open. It is possible that 3D sound could eliminate the psychological negatives of STT and TTS by creating a immersive environment which creates an intimate bond between the user and computer. 3D sound would in turn allow the user to speak naturally at a intimate level of voice which is "usually very close to our head and in a very low voice or even whisper" (Hall, 1996). Thus at an intimate level the user voice is not projected loudly and nor is the computer's as it places the sound closer to the users head. This level of intimacy and engagement through proximal control would eliminate most concerns of anxiety attributed to privacy and eavesdropping.

Virtual Environments

Another applicable area to 3D audio is it usage in virtual environments such as chat rooms. With 3D audio it is possible to create a virtual space with sound. This would create ambient and background noise of other people in the room, but only the direct conversation will be crystal clear. Additionally, people can take on a physical location in space as one person you are talking to appears to on your left, another person is to your right, with a friend approaching from behind you. Chat rooms can be further enhanced by eliciting the same proximal cues that occur in real world communication. Your friends in the chat rooms feel closer and reside in your intimate and personal space where other people feel as though they are in your social space. Additionally, talks could be given over the internet to feel like a virtual auditorium utilizing cues in our public space. Furthermore, people can follow social norms of non-verbal cues so when someone pays you a complement it feels as though they are moving closer to you. Additionally, online romances will benefit greatly as well.

Teleconferencing

Teleconferencing like virtual environments will also benefit from creating a 3D environment. Instead of just talking into space, our voice can be directed towards the person as they appear to be to our left, center, right, front, or behind. With 3D sound it is possible to create a phantom meeting where everyone feels like they are sitting at a table as oppose to being on the phone. Additionally, the conversations will be more salient as the sounds are richer, clearer, and more life like compared to traditional telephony.

Virtual Mapping

The last potential application I will talk about deals with expanding a computer screen beyond its physical dimensions. With 3D sound it should be able to put applications off into space, thus building a physical office in the user's mind. These cues could create a office by making it feel as though the mail box which receives email is off to your left, your fax is off to your right, etc. In this scenario sounds would be associated with these events, but the sounds would appear to becoming from different parts of the room. This concept could also be incorporated into networks to use the principles of social facilitation in office settings. Essentially, this could eventually be used in reverse of the solution for with Speech to Text and create a "big brother is watching you" effect in specific situations.

Another way 3D audio can be used is for expanding the desktop in the use of agents. Programs are increasingly being designed with personal agents and experts in mind. 3D sound could allow for a virtual room of experts which would allow it to feel as though certain experts were positioned around you. Furthermore, proximal rules could be applied to create a social distance which is more formal and giving more credibility to the agents.

While there are many implications, these are only a few of how 3D audio can effect voice interfaces and applications in the future. The next step is to move beyond the hypothetical implications of proximal control through 3D audio and to gather facts through a set of experiments.

REFERENCES

Carlile, Simon (1996) *Virtual Auditory Space: Generation and Applications* (Chapman and Hall, New York).

Handel, S. (1989) *Listening* (MIT Press, Cambridge, MA).

Maddox, Richard (1995) "Psychoacoustics and the home theater experience," *Home Theater Magazine*, April.

Mills, A.W. (1972) "Auditory localization," in J.V. Tobias, Ed., *Foundations of Modern Auditory Theory, Vol. II*, pp.303-348 (Academic Press, NY).

Wright, M. (1996) "3-D Audio: Above and behind you or moving low left to right," EDN, pp. 87-99 June.

WWW REFERENCES

Duda, Richard (1997) *3-D Audio for HCI*
www-engr.sjsu.edu/~knapp/HCIROD3D/3D_home.htm

Intel (1997) *Intel Realistic Sound Experience (3D RSX),*
www.intel.com/ial/rsx/wpaper.htm

(1997) *White Paper: Spatial Audio Technology,*
creative-ca.creaf.com/pressroom/info/papers/paper-spat-audio.html

Audio and Three Dimensional Sound Links
www.users.dircon.co.uk/~wareing/3daudio.htm

SRS Labs
www.srslabs.com