Virtualized Audio

Abel Ndabambi
Department of Computer Science
University of Cape Town
South Africa
+27 21 650 2663
andabamb@cs.uct.ac.za

Carl S. Schutte
Department of Computer Science
University of Cape Town
South Africa
+27 21 650 2663
cschutte@cs.uct.ac.za

Technical Report Number
CS04-25-00

ABSTRACT
In this paper, we describe a study of human auditory perception which focuses on sound localization and speech intelligibility. We present the results from experiments (n=4) which directly compared stereo headphones to 5.1 surround sound speakers. The results show that although localization ability is worse on headphones, it can be significantly improved by simulating free-field conditions. The results further show that speech intelligibility can be improved by using headphones and by separating speech and noise.

Categories and Subject Descriptors
H.5.5 [**Information Interfaces and Presentation**]: Sound and Music Computing – *Methodologies and techniques, modelling, systems.*

General Terms
Experimentation, Human Factors

Keywords
Audio, Localization, Sound, Speech Intelligibility

1. INTRODUCTION
For a long time, vision has been the dominant perceptual sense and the principal means for acquiring information. The introduction of digital sound within computers, and in particular the manipulation of sound in virtual space, has only recently come to the attention the general public. The synthesis and the manipulation of auditory space embodies new domains of experience that promise to change the way people think about sound.

Our ability to hear the world around us in three dimensions comes so naturally, that we almost take it for granted. Most of us have, at least, a basic idea of the physical mechanisms of human hearing, but few of us understand our spatial hearing abilities. This is largely due to textbooks overlooking this remarkable ability.

Admittedly, blame should not fall on the authors of these "incomplete" textbooks, as an understanding of these abilities, lies somewhere between Physics and Psychology. Psychoacoustics aims to quantify and explain these abilities. This hybrid area of research has helped us to understand the process by which out head and ears receive and encode all incoming signals, which are subsequently decoded and processed by the brain.

With current technology we are able to synthesize these 3-D sound processes. This has been of particular importance to virtual reality applications, where sound has been shown to be an increasingly important component [2].

This research aimed to investigate human auditory perception. The study focused on two important areas of audio research, namely sound localization and speech intelligibility. The sound localization aspect investigated a method of potentially simulating a more realistic virtual sound field, while the speech intelligibility aspect investigated the effects of angular separation of speech and noise. Both aspects compared stereo headphones with 5.1 surround sound speakers.

In section 2 we include a short background. Section 3 describes the experimental design, while the results are discussed in section 4. Finally, we conclude and offer suggestions for future work in section 5.

2. BACKGROUND
2.1 Binaural Hearing and Localization
Lord Rayleigh pioneered much of the work on the spatial properties of hearing. He found that low frequency sounds were more difficult to locate than high frequency sounds. According to Rayleigh's explanation, a sound coming from one side of the head produces a more intense sound in one ear than in the other ear, because the head casts a "sound shadow" for sounds of high frequency. This shadow effect is small for low-frequency sounds, because sound waves of long wave length diffuse around the head [8].

However, we are still able to localize low-frequency, just with slightly less accuracy than high-frequency ones. Rayleigh offered a second theory to explain low-frequency effects which states that a sound coming from one side strikes one ear before the other, and thus the sounds in the two ears will be slightly out of phase.

At frequencies below 1 kHz, localization is mainly due to an Interaural Time Difference (ITD) between sounds. Above 4 kHz, the accuracy of localization declines, with a high error rate around 3 kHz demonstrating that the two mechanisms do not overlap

appreciably. The pinnae aid in the localization of sounds above 5 kHz. They help alleviate reversal confusions, because they receive more efficiently from the front [9].

2.2 Head-Related Transfer Functions

Virtual reality applications typically concentrate on modelling an individual listener within a virtual acoustic space. In that case it is generally desirable to directly model the physical cues that the listener would hear if the sound source were actually located in a certain direction. This can be done by recording how a real source sounds when it is in a particular position. In order to capture pinna cues, the shadowing effect of the head, reflections from the shoulders, etc., it is necessary to make the recording inside the ear canal as close to the eardrum as possible. Such recordings can then be used to derive a head-related transfer function (HRTF), which is essentially a filter through which any sound can be processed.

A significant problem for the implementation of 3-D sound systems is the fact that spectral features of HRTFs differ among individuals. Hence, it makes sense to determine how localization of virtual sound sources can be degraded when listening through another set of pinnae. Fisher and Freedman [5] showed a significant decrease in azimuth localization accuracy when listening through artificial pinnae versus the subjects own pinnae. One aspect of localization that becomes obvious, especially with nonindividualized HRTFs, is the variation of performance between individuals.

2.3 Speech Intelligibility

In many social situations, listeners receive simultaneous sounds from different sources. Most people are able to perceptually "tune out" the interfering or masking noises that emanate from various directions, focusing instead on signals of interest. This ability to recognize or understand speech in the presence of masking or competing noise, defined as the "cocktail party effect" by Cherry, depends upon several complex variables [3]. Acoustic parameters, environmental variables, and contextual variables contribute to speech intelligibility in both monaural and binaural listening conditions.

The contribution of monaural cues to speech intelligibility appears to greater than binaural cues [4]. However, the advantages of spatial hearing provided by binaural listening emerge in adverse conditions, such as low speech-to-noise ration (SNR), reverberation or a combination of these conditions [4, 10]. Previous experiments have shown that in monaural listening conditions, the following factors contribute to speech intelligibility: signal or sound source characteristics, masker characteristics, SNR, and redundancy of speech message.

Even though most people hear very well with headphones, free-field binaural hearing offers several advantages, including: localization of sound in space, sound separation, and enhances intelligibility in noise and reverberation [10]. Another possible contributor to increased speech intelligibility in free-field listening, as opposed to monaural headphone listening, is the central auditory system's ability to suppress noise internally, in some binaural listening conditions, based on interaural differences. This ability, called masking level difference, is predominately a laboratory phenomenon, but it indicates that the auditory system can internally improve the SNR if the interaural difference for speech and noise are different. The same underlying concept may apply to conversations at a party [10]. It has also been shown that angular separation of speech and masking noise can improve speech intelligibility [4].

3. EXPERIMENTAL DESIGN

3.1 Localization

As mentioned earlier, the main aims of the localization experiments were to investigate the differences between stereo headphones and 5.1 surround sound speakers. These headphones, however, would be presenting the listener with a dynamically updated free-field sound simulation. We also aimed to investigate the differences between localizing pure noise and speech in the presence of noise, when combined with our free-field simulation technique.

3.1.1 Design

In order to determine the feasibility of our hypotheses, it was necessary to perform experiments which would provide us with sufficient quantitative data, so that further statistical analysis might be conducted. The speaker setup and virtual sound source were identified as independent variables, while the listener's localization accuracy was identified as a dependent variable.

To gather the results we required, it was decided that subjects should use a pointing technique, where the results would be calculated from a motion tracker when the subject indicated they had faced the source. Based upon previous research, it was decided that we would not test elevation and distance perception. This decision was made due to the larger degree of error observed when listeners estimate the attributes.

3.1.2 Subjects

Two adults served as paid volunteers in the study (ages 20-22; 1 male, 1 female). Although we did not conduct audiometric evaluations, we screened subjects orally with questions directed towards the following issues: noticeable overall hearing loss, noticeable differential hearing loss, recent exposure to loud noise (e.g., amplified music, motorcycle), and medical history. The use of oral reporting is not unusual in localization studies; other localization studies that have used oral screening without audiometric screening include Noble [6], Asano, Suzuki, and Sone [1], and Perrott, Sadralodabai, Saberi, and Strybel [7]. All subjects completed a training block in order to acquaint them with the procedure.

3.1.3 Stimuli

The stimuli in the experiments consisted either of speech in combination with four broadband noise bursts, solely four broadband noise bursts. The broadband noise bursts were rectangularly gated to 500 ms; these noise stimuli were generated with a computer running MATLAB, and then output through the sound card at 44.1 kHz sampling rate to the audio input of the amplifier.

3.1.4 Procedure

The experiment was conducted with listeners located in a sound-treated listening room. Prior to the start of each trial of the

experiment, the listener was asked to turn to face directly at the computer and press the response button. This repose was used to "zero" the motion tracker by assigning that location to 0° azimuth. The first session output the stimuli to the surround sound speakers, while the second session output the stimuli to the headphones. Both sessions consisted of two blocks; the initial block presented the listeners with solely broadband noise bursts, while the second block presented listeners with speech in combination with broadband noise bursts.

For all blocks, the stimulus was randomly presented at one of 11 azimuth locations in the front right quadrant of the horizontal plane (spaced 9° apart), and the listener was asked to respond by turning to face directly at the apparent location of the stimulus and press the response button. Then the listener turned back to face directly at the computer to zero the motion tracker for the next trial, and the computer was also used to provide visual feedback about the location of the target stimulus, the location of the response, and the angular error between these two locations.

Each experimental session consisted of 264 trials. The first 132 trials of each session were conducted with broadband noise bursts, while the last 132 trials were conducted with speech in combination with broadband noise bursts. The trials consisted of the stimulus being repeated 12 times for each azimuth position. At the end of the block, were not given any information regarding their mean azimuth error of all the trials in that session. Both of the subjects participated in 2 of these experimental sessions. Thus, each subject participated in 528 trials.

3.2 Speech Intelligibility

The purpose of the experiments, were to investigate speech intelligibility by comparing headphones with surround sound speakers.

3.2.1 Design

The experiments involved playing the speech (target sound) and occluding it with broadband noise, through both headphones and surround sound speaker; then asking the subject to extract meaning from the speech.

3.2.2 Subjects

Two paid volunteers, naïve to the purpose of the experiments, served as subjects (ages 21-27; 1 male, 1 female). Both subjects were students with normal hearing. The subjects were made to feel comfortable and relaxed by giving each of them a short verbal introduction, orientation, and training block. The results of the training blocks were not taken into consideration when doing data analysis. The training blocks also server to determine the SNR that would be suitable for conducting the experiments. The SNR was determined by steadily reducing it from 1.0, until subjects could not score above 70%. The SNR that was found to be suitable was 0.9 and it was used throughout the experiments.

3.2.3 Stimuli

28 pre-recorded wave files were used as the speech or target sound, while the broadband noise served the purpose of occluding the speech. Each of the target sounds gave the subject an instruction to click on a combination of colour and number. For example, Blue5.wav contained the message: "Ready Baron, go to Blue 5 now." This instructed the subject to click on the "Blue" and "5" buttons.

3.2.4 Procedure

The target sounds were randomly selected and presented together with the masking noise to the subject, through both headphones and surround sound speakers. The subject then responded by clicking on the appropriate combination of buttons; with the system automatically capturing their response and processing it accordingly. The broadband noise always came from 0°, while the target sound came from four different azimuth positions (0°, 15°, 30°, and 45°) in random order with normal distribution.

Experiments were organized into 1 hour sessions of 4 blocks each; with each block consisting of 112 trials. The experiments were conducted over a period of 2 days, resulting in total 1792 trials; 896 trials with headphones and 896 trials with surround sound speakers.

4. RESULTS

4.1 Localization

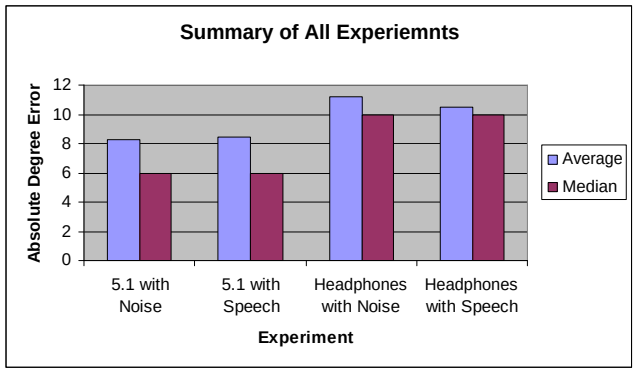

Figure 1. Data from localization experiments.

For the surround sound speaker experiments, an average absolute error value of 8° and a median value of 6° were observed under both conditions.

For the headphones experiment, that presented the listener with solely broadband noise bursts, an average absolute error value of 11° and a median value of 10° were observed. Similarly, for the headphones experiment, that presented the listener with speech in the presence of broadband noise bursts, an average absolute error value of 10° and also a median value of 10° were observed.

4.2 Localization

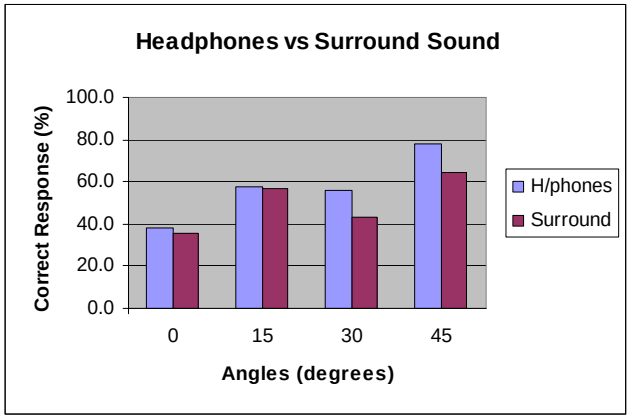

Figure 2. Data from speech intelligibility experiments.

For the headphones experiments we observed the following correct response rates at predefined azimuth positions: 37.9% at 0°, 58.0% at 15°, 56.3% at 30°, and 78.1% at 45°. We further observed an average and median correct response rate of 57% for the headphones experiments.

Likewise, for the surround sound speakers experiments, the following correct response rates were observed: 35.3% at 0°, 57.1% at 15°, 42.9% at 30°, and 64.7% at 45°. We also observed that the surround sound experiments produced an average and median correct response rate of 50%.

4.3 Discussion

With regards to localization, we can infer that the surround sound experiments yielded almost identical error value under both conditions, showing no benefit in speech over noise. Likewise, both headphones experiments yielded very similar error values, again showing no benefit in speech over noise. In comparison with the surround sound experiments, the headphones experiments showed higher error values. However, we still feel that these error values are particularly low for localization with stereo headphones.

With reference to the speech intelligibility experiments, we can deduce that both listeners performed better with headphones than with surround sound. The results further showed, that the greater the degree of angular separation, the better the listener's performance. However, it is interesting to note that the listener's performance at the 30° azimuth position was slightly lower than at the 15° azimuth position.

5. CONCLUSIONS

Based upon our results from the localization experiments, we can conclude that although localization accuracy was worse on stereo headphones, it can be significantly improved by employing our technique, whereby we dynamically simulated free-field listening conditions. Furthermore, we were able to conclude that presenting listeners with speech, in the presence of noise, provided no significant localization benefit. Based upon our results from the speech intelligibility experiments, we could conclude that speech intelligibility can be improved by using headphones and by separating speech and noise.

For future work, we recommend that the new localization technique be more thoroughly investigated. This could be done by increasing the sample size and by comparing the dynamically updated headphones with normal headphones. The uncharacteristic result produced by the speech intelligibility experiments at 30° azimuth, warrants further investigation. We propose that more than four azimuth positions be used and the sample size be increased.

6. ACKNOWLEDGMENTS

We wish to thank our supervisors, Prof. E. H. Blake and Mr. J. Verwey, whose constant encouragement and support led to the success of this study.

7. REFERENCES

[1] Asano, F., Suzuki, Y., & Sone, T. Role of spectral cues in median plane localization. *J. Acoust. Soc. Am., 88* (1990), 159-168.

[2] Begault, D. R. *3-D sound for virtual reality and multimedia*, Academic Press Professional, Inc., San Diego, CA, 1994.

[3] Cherry, E. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Amer. 25* (1953), 975-979.

[4] Ericson, M. A. and McKinley R. L. The Intelligibility of Multiple Talkers Separated Spatially in Noise, in *Binaural and Spatial Hearing in Real and Virtual Environments*, Gilkey, R.H. and Anderson, T.A. (eds.), L. Erlbaum Associates, Mahwah, NJ, 1997, 701-724.

[5] Fisher, H., and Freedman, S. J. The role of the pinnae in auditory localization. *J. Auditory Research, 8* (1968), 15-26.

[6] Noble, W. Auditory localization in the vertical plane: Accuracy and constraint on bodily movement. *J. Acoust. Soc. Amer. 82* (1987), 1631-1636.

[7] Perrot, D. R., Sadralodabai, T., Saberi, K., and Strybel, T. Z. Aurally aided visual search in the central visual field: Effects of visual load and visual enhancement of the target. *Human Factors, 33* (1991), 389-400.

[8] Lord Rayleigh (Strutt, J. W.) *The Theory of Sound*, Vol. 1, 2nd ed. London: Macmillan. Reprinted by Dover, New York, 1945.

[9] Rossing, T. D., Moore, F. R., and Wheeler, P. A. *The Science of Sound* Addison-Wesley, Reading, MA, 2002.

[10] Vause, N. L., and Grantham, D. W. Speech Intelligibility in Adverse Conditions in Recorded Virtual Auditory Environments. In *Proceedings of the International Conference on Auditory Display '98* (University of Glasgow, UK, November 1-4, 1998).