



Audio Engineering Society Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

This convention paper has been reproduced from the author's advance manuscript, without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. Additional papers may be obtained by sending request and remittance to Audio Engineering Society, 60 East 42nd Street, New York, New York 10165-2520, USA; also see www.aes.org. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Telescopic Spatial Radio

Norman P. Jouppi¹, and Subu Iyer¹

¹Hewlett Packard, 1501 Page Mill Rd., Palo Alto, CA 94304 USA

Correspondence should be addressed to Norman P. Jouppi (norm.jouppi@hp.com)

ABSTRACT

We have developed a system we call Telescopic Spatial Radio (TSR). This system transforms monaural transmissions from geographically distributed speakers into a spatial audio presentation using binaural techniques which preserve the actual physical angles between participants. TSR instantly augments the user's situational awareness with the headings of the speaking users. The system leverages orientation measuring, location tracking, and signal processing capabilities that are rapidly decreasing in cost. TSR has many potential applications ranging from emergency and aviation communication to a richer consumer experience. We have developed a prototype system using laptop computers, GPS, and electronic compasses. The system allows users to select HRTFs from a library, and operates over a computer network.

1. INTRODUCTION

In many applications involving remote communication between people, situational awareness of the relative position of the participants can vary from a pleasant feature to providing critical information. For example in search and rescue applications, it can be critically important to know the relative heading to other team members as they speak. In social applications, spatial audio may merely create a richer and more enjoyable user experience. When people are within natural speaking distances of each other, the relative position of speakers is immediately conveyed by the perceived direction of their voice. Providing the same directional advantages enjoyed by collocated speakers to geographically distributed re-

mote speakers could significantly enhance their situational awareness.

Remote communication systems usually need to support multiple actively speaking users. However, current radio and walkie-talkie systems only support a single active talker at a time per radio channel. (In the case of more than one active speaker, all the speakers or all the speakers except the loudest may be garbled.) In contrast, a system modeled on collocated speakers would enable the users of the system to perceive the voices of geographically distributed simultaneous speakers as coming from different directions. If the angle between the speakers was significantly different for a given listener, the listener could selectively attend to the speaker of their choice

using human perceptual abilities commonly known as the “cocktail party effect”.

We are working to provide advantages of colocated speakers to distributed remote speakers. We call the system with these capabilities Telescopic Spatial Radio (TSR), since it maintains the relative orientation of speakers and listeners, but at seemingly much reduced distances. This is similar to the use of telescopes to make objects at given headings appear closer.

2. SYSTEM GOALS

In an ideal Telescopic Spatial Radio (TSR) system, geographically distributed people could communicate with each other using devices similar to radios or walkie-talkies. However, these radios would be augmented with headphone/mic headsets, signal processing capabilities, and position and orientation tracking. They could be connected either by a wireless computer network or by one or more conventional radio frequencies. The signal processing capabilities would be used to spatially encode other people’s voice communication for output over the user’s headphones using head-related transfer functions (HRTFs). Based on the location of the speaking person, the location of the user, and the user’s orientation (determined by a low-cost digital compass integrated into the headphones), the perception of the direction of the speaker could be immediately conveyed to the user (see Figure 1). For example, if the user is facing true north, and the speaker is a mile away from the user to their west (i.e., left), the speaker’s voice will be processed to appear to be coming from the left of the user. If the user then turns their head while that speaker continues to talk, the voice of the speaker will appear to be fixed in space relative to the listening user. Similarly, if the speaker is at the top of a hill and the user is in a valley, the speaker’s voice should be perceived as coming from above by the user at the elevation angle existing between the users.

3. PROTOTYPE IMPLEMENTATION

We have implemented a prototype TSR system running on laptop computers augmented with GPS, headphones, and 3-axis electronic compasses. The system allows users to select HRTFs from a library[16], and uses the HRTFs to produce motion-tracked binaural sound[1] for the users. The system

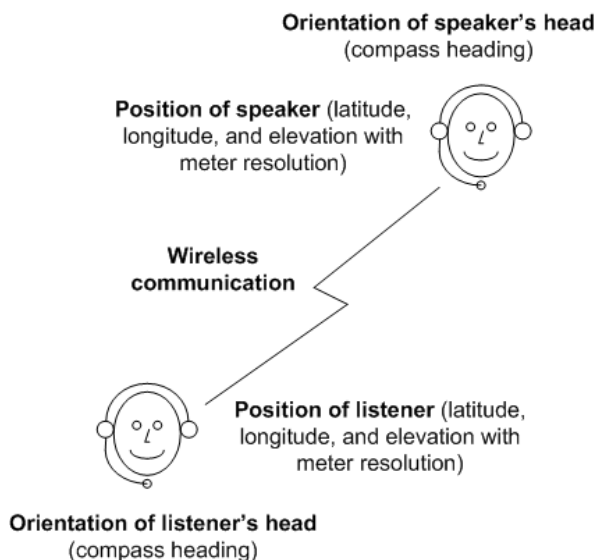


Fig. 1: Overview of system operation. Each speaker’s audio channel is augmented with their position and orientation information. This is combined with the position of each listener and the orientation of their head to convey the heading of the speaker via motion-tracked binaural sound.

is currently designed for sample rates of 22.05 KHz or 44.1KHz. We use Microsoft’s Audio Compression Manager (ACM) for transparent run-time audio compression and decompression[12]. Microsoft ACM uses a library of external codecs to convert audio data from one format to the other. This gives us the ability to select different codecs depending on bandwidth and latency requirements. Currently up to four active speakers are supported over a computer network.

We reduce network bandwidth by using multicasting[17] for sending audio from speaking users. In multicasting, the bandwidth requirement for the speaker’s system is the same regardless of the number of receivers. The speaker’s system sends streaming audio to a particular IP-Multicast address to which all the remote machines subscribe. The substantial bandwidth savings from multicasting also helps to reduce network congestion and server load, and it enables the participation of potentially thousands of remote listeners.

Although our prototype TSR system is based on a computer network, it could also be supported with analog radio systems capable of receiving several radio channels simultaneously. (Note that digital radio systems employing orthogonal frequency-division multiplexing already receive several frequencies simultaneously.) In a system utilizing a band of analog radio channels, speakers could be assigned to a particular frequency based on whichever channels were not in use and had the least noise. The position of the speaker (e.g., GPS coordinates) could be repeatedly broadcast on a low bitrate orthogonal channel (<1Kbs). This is similar to the broadcasting of digital data in the third harmonic of the stereo subcarrier in the Radio Data System[10].

Our system uses GPS receivers to determine the location of each user, providing their latitude, longitude, and elevation. We use a commercially-available GPS receiver in a sub-PCMCIA form factor with an external antenna. GPS receivers are being rapidly reduced in size and cost to support widespread adoption in cellphones in support of 911 location awareness. The target cost for GPS receivers in volume is two dollars. The major limitation of GPS for our application is that it does not work indoors. For indoor applications other position sensing techniques or a combination of GPS and inertial navigation techniques would be required.

We also need to know the orientation of the user's head relative to the speaker. We currently obtain the orientation from a 3-axis digital compasses. Electronic compasses are also rapidly decreasing in cost and size; 3-axis compasses are approaching ten dollars in volume. A 3-axis compass is required because the user may tilt their head as well as rotate it. A 3-axis compass first determines the inclination of the compass, and then uses this to process the output of magnetic sensors in X, Y, and Z orientations into a compass heading. The compass should be placed at the top of the headphone band, since this reduces and balances the magnetic interference from the transducers in the headphone's speakers. The digital compass used in the prototype system is shown in Figure 2.

Convolving a single speaker with HRTFs using a 1.8GHz laptop only uses less than 10% of the processor for 22.05KHz sample rates, even though the

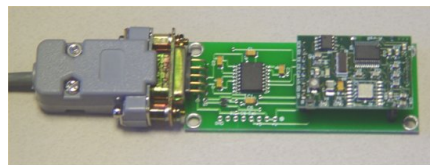


Fig. 2: The digital compass used in the prototype. The actual compass is the small daughter board to the right; the scale is given by the serial connector on the left.



Fig. 3: Remote prototype setup including user interface. (The compass and GPS plug into the laptop.)

HRTF processing is performed in C++ code without using specialized instructions for higher floating-point performance. Embedded signal processing to convert incoming monaural radio transmissions into binaural presentation for the user should require processor resources costing only a few dollars.

4. IMPLEMENTATION ISSUES

There are many design alternatives possible in a TSR system. These include methods of HRTF selection and adjustment, HRTF storage, simulation of reflections and distance, control of the telescopic zoom settings, and volume level control. Other design issues include methods for graceful degradation with loss of position or orientation data, opportunities for virtual positioning, compatibility with legacy systems, and the networking topology used.

4.1. HRTF Fitting

In our prototype we allow the user to select the HRTF with the best match from the HRTF database[16]. One common issue with the use of generic HRTFs is elevation mismatch due to pinnae differences. To null-out elevation mismatches we have added an elevation slider to the graphical user interface on our prototype. This merely shifts the elevation index into the HRTF database by the amount specified by the slider.

In a practical implementation of TSR, users would need the ability to select between multiple HRTFs to find a good fit for their own HRTF. Recent work has shown that a relatively small number of HRTFs can provide a good fit for most users[14]. Seeber and Fastl found that a selection among the most common 5 HRTFs provided a good fit for approximately 85% of the test subjects. In practice, users should be able to try out HRTFs in order of the most popular to the least popular, to minimize the number of evaluations required before being able to start using the system. The ability to enter relevant morphological information (like the user's age) could also be helpful in improving the efficiency of the fitting process.

In some cases a more complicated fitting process over a larger database might be required, or in extreme cases users might want to be able to supply their own HRTF. In this case HRTFs could be downloaded into the radio either by short range wireless radio (such as Bluetooth), infrared, or by the use of a flash memory card.

Finally, in cases where a proper fit could not be achieved due to time or other constraints, the system should support the disabling of spatial audio presentation. This would also be an important accommodation for people who had a significant hearing impairment, such as being deaf in one ear.

4.2. HRTF Storage Requirements

The storage requirements for one complete HRTF of 200 filter coefficients, 50 azimuths, and 25 elevations for both ears[16] requires 2 megabytes. With continued scaling of Moore's Law this has recently become a relatively small amount of data. For example, flash memory cards containing 512MB are currently available for 77 dollars in single retail quantities, or about 15 cents per megabyte. Furthermore,

HRTF databases contain a fair amount of redundancy, so they can be significantly reduced in size through data compression. We have found compression down to 33% of the original data set size are possible using ordinary compression techniques such as Lempel-Ziv (i.e., Unix compress or WinZip). (Compression techniques tailored for the HRTF databases should be able to achieve even higher compression rates.) Nevertheless, even with generic compression techniques, the storage to contain 5 full HRTFs in compressed form would only require less than a dollar of flash storage at today's retail prices. The ability to load HRTFs via a wireless connection or a flash memory card slot could likewise be provided at modest cost. Finally, the size of the HRTF database could also depend on the extent that HRTF interpolation techniques were used[11].

4.3. Reflections

In order to provide the most effective user experience, even if users were in outdoor settings, the presence of the system's users in a shared physical room should be simulated. This is because reflections aid in creating a perception of spaciousness and creating the perception of direction. Basic reflections could include listener torso reflections and a first reflection from a simulated floor. More extensive modeling of reverberation could also be used to help indicate relative distances[15].

4.4. Simulation of Distance

It is also possible to convey a perception of relative distance by signal processing of the audio waveforms presented to the user[18]; however people are not as accurate in perceiving distances as they are in perceiving directions. This processing would modify the timbre and volume of the waveform. Obviously there is usually no point in simulating the actual distance, since speakers may be many miles apart. Also, there are limitations on how much a speaker's voice can be attenuated without reducing its intelligibility. However, within limited ranges, some perception of relative distance could be simulated. The distance to the furthest speaker could be ratioed to a maximum distance without resulting in unacceptable intelligibility degradation, and speakers closer than this could be presented as being closer by the same ratio. For example if the maximum simulated distance was 25 feet, and the distance to the furthest user was 25 miles, all speakers could be ren-

dered to be at a perceived distance of one foot from the listener per each actual mile. Other, non-linear mappings may be used as well.

4.5. Telescopic Zoom: Manual and Automatic

It is also possible to control the magnification of the sounds from distant participants either manually (similar to a zoom lens on a telescope) or automatically.

In the manual case, the user could be given a control which sets the falloff of sound with distance. By varying this control the user could restrict their hearing to varying radii around their position. This is useful in cases where the system may be in use by many teams of people over a large area, for example fire fighters in a large city such as New York.

In the automatic case, the distance functions could vary dynamically depending on the number of active speakers. In this scenario, the sound would fall off slowly with distance if there were few active talkers, but if there were many active talkers, sounds would be processed to falloff more rapidly with distance. The falloff rate would be chosen so that the number of active speakers heard by the user would not exceed the number of conversations discernible by the listener using the cocktail party effect. Hence the falloff function could also be a function of the relative position and distances to the speakers. For example, if many equal-loudness speakers were spread optimally around a listener[13, 4] the falloff with distance could be reduced. An automatically zooming system would have the disadvantage that remote speakers would appear and disappear based on the presence of close speakers, especially if they had the same heading. This might generate some confusion on the part of listeners, and should be investigated experimentally.

The effectiveness of both manual and automatic zoom control would also be a function of level control, as discussed in the next section.

4.6. Volume Level Control

In many radio systems, the goal is to present both close and distant speakers with the same loudness. This is accomplished through the use of automatic gain control at one or more points in the system. However, in a TSR system, depending on the dynamic range of the communication channel and the

ambient noise levels experienced by the users, volume levels closer to calibrated signal levels could be used. In this case, speakers who were speaking very loudly would be perceived as speaking more loudly than those who were speaking softly. Thus even at distances beyond those normally intelligible based on the zoom level, loud discourses could be heard and understood. The level control could be designed to vary automatically from calibrated to compressed based on the ambient noise level.

The volume of the talker could also be made dependent on the orientation of the talker[7], similar to the case where the talker and listener are physically co-located. In this way a talker could selectively direct their voice to different listeners depending on the orientation of their head or via user interface controls. The directivity is also obviously a function of the virtual room reverberation model being used. Since the talker is also a listener, they already have head orientation sensing. If the talker loudness was also made dependent on the talker's head orientation, the orientation could also be sent on the same orthogonal data channel as the talker's position data. This would add an insignificant amount of additional bandwidth (only a few dozen bits/sec could suffice). Since the directivity of human talkers is more uniform among different subjects than HRTFs, only a single data set specifying directivity with azimuth and elevation would be required. Even if this was frequency-dependent, this would require storage less than or equal to an HRTF.

4.7. Graceful Degradation During Tracking Loss

Because the spatial audio presentation in TSR is dependent on position information from the speaker and listeners as well as the orientation of the listeners, loss of this information has implications for the system operation.

4.7.1. Loss of Position

If a user went indoors (in a location without indoor GPS repeaters), loss of position information could occur. However, in order to measure the user's head orientation with a 3-axis compass, X and Y accelerations must be measured. If Z accelerations are measured as well, and all three accelerations are integrated, this forms the basis of an inertial navigation system. This could be used to predict a listener's position when out of contact with GPS (or other primary position tracking system). Because

inertial navigation has drift and other errors that can accumulate over time, the precision of the position estimate will degrade with time. The expected error bounds on position can also be broadcast on the orthogonal data channel with very little additional bandwidth. If the error becomes significant relative to the position of a listener (e.g., it could cause the speaker to be either to the right or left of the listener), then spatial presentation of that speaker's voice could be disabled for that listener. Spatial audio presentation would also need to be disabled in cases where the system had obtained no primary position information since being powered on. In this case the position uncertainty would be set to its maximum value.

4.7.2. Loss of Orientation

The loss of orientation data is less likely than the loss of position information, since compasses also work indoors. Compass headings can be affected by large magnets, electric currents, or ferrous structures. This could cause errors in the orientation data, however most environments should be fine. In extreme environments magnetic sensing could be augmented with angular rate gyroscopes, similar to the augmentation of GPS with inertial navigation as described above. Good quality electronic gyroscopes can now be obtained for under ten dollars in volume.

4.8. Virtual Positioning

Listeners and speakers would not need to be limited to their actual physical position or orientation. For example, a fire chief might want to participate in operations from the auditory vantage point of being in the thick of things, rather than being off to one side. Thus our prototype allows a user to specify either an absolute virtual position or a virtual position relative to their actual physical position.

You could also imagine a situation on a battlefield where a remote senior commander would want the vantage point of a field commander or a team. In this case the virtual position could be set to automatically "tag along" with a particular speaker, or to automatically be centered in the middle of recent speakers.

4.9. Compatibility with Legacy Systems

In a digital system, all users are likely to be using compatible systems. However if TSR capabilities are overlaid on traditional analog communication

channels (using techniques similar to those utilized by the Radio Data System), some users with older equipment may not transmit their position (and optionally orientation) data. In other cases, if there were large amounts of radio interference, transmission over long distances, or obstacles to radio transmission, analog voice signals might be intelligible but the orthogonal data transmission might be lost. In both these cases the speech of talkers would not be able to be spatialized, but would be presented equally to both ears of the listener.

4.10. Network Topology

The TSR devices could be connected by a wireless network in either a distributed or centralized fashion (see Figure 4). In the distributed arrangement, each TSR device can either send its user's voice information individually to the other TSR device or as a group via multicasting. In the centralized arrangement, each TSR device can communicate with a server which aggregates the voice traffic from each speaking user and sends the combined data back to all the users. The server configuration can be especially useful in cases where users may be out of radio contact with each other, but are all connected to a transponder on high ground (including satellite systems).

In the server arrangement, multiple audio streams from multiple simultaneous streams are combined on the server and only a pair of audio channels is transmitted to each user. In the distributed scenario, each user must receive audio from every active speaker, and process it based on HRTFs and combine it locally. This requires more local compute power than the server approach, especially in the case of multiple simultaneous speakers, but eliminates the need for a centralized server. Multicasting (for computer networks) or broadcasting (for radio systems) can be used in the distributed case so that the audio of each speaking user needs to be transmitted only once - otherwise the network transmissions per speaking user could go up as the number of active listeners.

5. APPLICATIONS

Telescopic Spatial Radio's ability to augment the user's situational awareness with the headings of the speaking users can be valuable in many applications. For example, in aviation, the direction of speakers can be crucial. With TSR when an airplane has

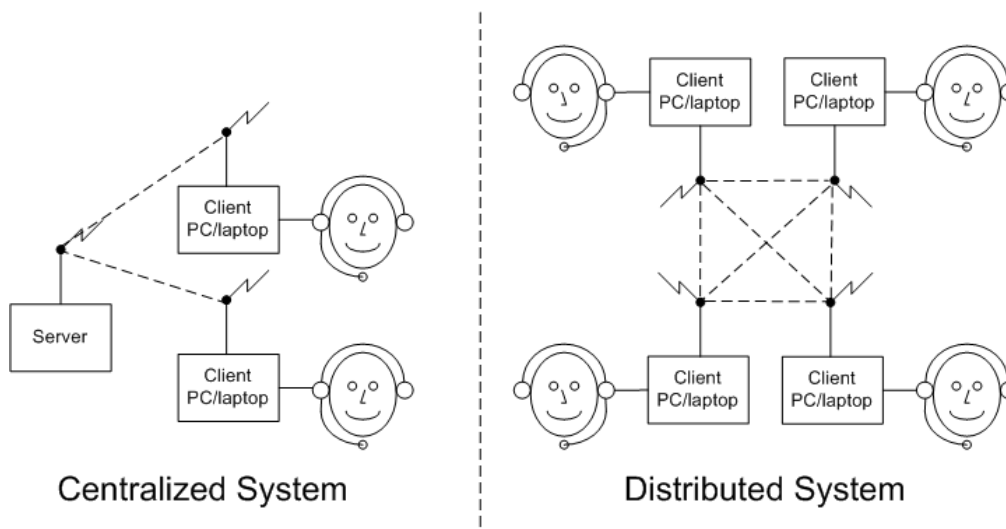


Fig. 4: Two possible prototype network architectures, using either centralized or distributed communication. Our prototype uses distributed communication with multicasting.

passed over a control tower the voice of the controller would be perceived as passing below and behind the pilot. Pilots speaking from other planes around the pilot could also be heard speaking from the directions of their airplanes. This technology would also have applications in military communications and could reduce the incidence of friendly fire.

Although not life-critical, the ability to enjoy directional spatial audio at a distance for a modest additional cost could enhance the rich media experience of users in many different communications applications. For example, many cellphones may be augmented with GPS capabilities for provisioning 911 services. If electronic compasses are also included, then cellphones of the future could have the tracking capabilities required for implementation of TSR. Thus one might hope that TSR could aid people in meeting each other in crowded locations, or allow multiple distributed people to chat with each other.

However, in the case of a cellphone the orientation tracking function is more likely to be implemented in a cellphone handset instead of in lightweight earbud headphones. In this scenario, motion-tracked binaural sound could not be provided based on the user's head, but only on the approximation of their median head orientation based on an expected angular off-

set from their cellphone handset mounted on their belt or in a pocket. Thus, users would only get good orientation spatialization while orienting their head forward. Of course, such uses would require much better voice quality, but that is an issue with many non-technical considerations.

6. RELATED WORK

Recently there has been much work in spatial auditory displays for presenting information using spatial displays and for distinguishing among many simultaneous talkers[8, 3]. However, the focus of this prior work has been on maximizing intelligibility of complex audio scenes by varying the position of the sources. In contrast, the focus of TSR is on providing a physically accurate spatialization of multiple talkers based on the actual physical locations and orientations of the participants, in order to convey heading and distance information along with a talker's voice.

7. SUMMARY

Telescopic Spatial Radio can naturally convey multiple speakers' headings to listeners over large distances. TSR can also support discrimination among simultaneously speaking users at different headings via human perceptual abilities (i.e., the "cocktail

party effect”). The orientation and position tracking technologies required to implement TSR are available today and are rapidly decreasing in cost. Furthermore, signal processing to convert incoming monaural radio transmissions into binaural presentation requires only modest computational resources. We believe Telescopic Spatial Radio could provide significant value in many applications, and is worthy of further research and development.

ACKNOWLEDGEMENTS

The authors would like to thank Jacob Augustine, Shivarama Rao Kokrady, Deepa Kuttipparambil, and Naveen Belkale for their help with the software.

8. REFERENCES

- [1] V. R. Algazi, R. Duda, and D. Thompson. Motion-Tracked Binaural Sound. In *Proc. of the 116th Audio Engineering Society Convention*, May 2004.
- [2] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 2nd edition, 1997.
- [3] D. S. Brungart, M. A. Ericson, and B. D. Simpson. Design Considerations for Improving the Effectiveness of Multitalker Speech Displays. In *Proc. of the International Conference on Auditory Display*, July 2002.
- [4] D. S. Brungart and B. D. Simpson. Distance-Based Speech Segregation in Near-Field Virtual Audio Displays. In *Proc. of the International Conference on Auditory Display*, July 2001.
- [5] C. Cheng and G. Wakefield. Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space. *Journal of the Audio Engineering Society*, 49(4):231–249, April 2001.
- [6] C. Cherry. Some Experiments on the Reception of Speech with One and Two Ears. In *Journal of the Acoustical Society of America*, pages 975–979, 1953.
- [7] W. T. Chu and A. C. C. Warnock. Detailed Directivity of Sound Fields Around Human Talkers. *National Research Council Canada Technical Report*, IRC-RR(104), December 2002.
- [8] K. Crispian and T. Ehrenberg. Evaluation of the “Cocktail Party Effect” for Multiple Speech Stimuli within a Spatial Auditory Display. *Journal of the Audio Engineering Society*, 43(11):932–941, November 1995.
- [9] R. H. Gilkey and T. R. Anderson. *Binaural and Spatial Hearing in Real and Virtual Environments*. Lawrence Erlbaum Associates, 1997.
- [10] D. Kopitz and B. Marks. *RDS: The Radio Data System*. Artech House Publishers, 1999.
- [11] M. Matsumoto, S. Yamanaka, and M. Tohyama. Effect of Arrival Time Correction on the Accuracy of Binaural Impulse Response Interpolation. *Journal of the Audio Engineering Society*, 52(1):56–61, January 2004.
- [12] Microsoft. Microsoft Audio Compression Manager. http://msdn.microsoft.com/library/en-us/multimed/htm/_win32_audio_compression_manager.asp.
- [13] W. T. Nelson, R. S. Bolia, M. A. Ericson, and R. L. McKinley. Spatial Audio Displays for Speech Communications: A Comparison of Free Field and Virtual Acoustic Environments. In *Proc. of the Human Factors and Ergonomics Society 43rd Annual Meeting*, 1999.
- [14] B. U. Seeber and H. Fastl. Subjective Selection of Non-Individual Head-Related Transfer Functions. In *Proc. of the International Conference on Auditory Display*, July 2003.
- [15] B. Shinn-Cunningham. Speech Intelligibility, Spatial Unmasking, and Realism in Reverberant Spatial Auditory Displays. In *Proc. of the International Conference on Auditory Display*, July 2002.
- [16] C. U.C. Davis. The CIPIC HRTF Database. http://interface.cipic.ucdavis.edu/CIL.html/CIL_HRTF_database.htm.
- [17] R. Wittmann and M. Zitterbart. *Multicast Communication: Protocols and Applications*. Morgan Kaufmann, 2001.
- [18] P. Zahorik. Auditory Display of Sound Source Distance. In *Proc. of the International Conference on Auditory Display*, July 2002.