



From Cocktail Parties to Conference Rooms:  
Using Human Audition to Improve Collaborative Work

---

KMDI 1001 Assignment One

Christopher Collins  
991 783 183  
Department of Computer Science, University of Toronto

May 12, 2006

## **Abstract**

Spatialized audio is a technology of increasing importance as interaction using audio in multiuser chatspaces and for collaborative work becomes more commonplace. Users are turning to audio as a modality of interaction online, and traditional forms of teleconferencing are evolving into telepresence — interactive environments that make use of as much perceptual information as possible. A review of relevant literature shows two main areas of human auditory function that hold promise for enhanced audio conferencing — the ability to detect spatial location and the “cocktail party effect”. Recent research shows the memory, comprehension, and enjoyment of audio conference participants can be enhanced by adding a spatial cues. Also, human listeners have the ability to monitor more than one sound channel at a time, and this ability is increased with spatial separation of sound sources. Production technologies range from those experienced by a single listener with headphones to those experienced by multiple listeners in a specially equipped room. This work explores how detailed knowledge of human audition is being exploited to enhance the experience and productivity of audio conferencing in computer supported collaborative work environments.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>The Cocktail Party Effect</b>	<b>2</b>
<b>3</b>	<b>Spatial Audio</b>	<b>5</b>
3.1	Perceptual Synthesis . . . . .	5
3.2	Sound Field Simulation . . . . .	7
3.3	Comparative Experiments . . . . .	8
3.4	Memory and Comprehension Experiments . . . . .	9
<b>4</b>	<b>Spatial Audio Conferencing Applications</b>	<b>11</b>
<b>5</b>	<b>Conclusion</b>	<b>13</b>

## 1 Introduction

Increases in bandwidth coupled with decreasing technology costs have resulted in a surge in the use of computers to support collaborative work between participants possibly remote from one another. Both in specialized facilities and directly through a standard desktop computer, people around the world are collaborating on a daily basis through video and audio conferencing. As the limits of bandwidth decrease, the focus of audio conferencing research is moving away from quality of service and synchronicity concerns toward ways in which the experience of *telepresence* can be enhanced to provide a richer, more productive medium (Baldis, 2001).

In a face-to-face meeting with multiple participants, there are many ways in which conferees can interact. These include visual signals (*e.g.*, facial expressions, body language) and audio signals (*e.g.*, timbre and intensity of voice). Using only on the audio modality, humans can tell the relative location of conference participants, identify speakers, and focus on one speaker while “tuning out” others (using selective listening and auditory scene analysis). Participants can engage in side conversations while passively monitoring the main discussion and remaining attentive to key words (the “cocktail party effect” (Handel, 1989)).

In a traditional audio conference (or video conference), the voices of multiple participants come from a single audio stream, delivered from a set of monophonic speakers, creating an unnatural interaction environment. The problems of this audio setup were first identified in the 1950s by air traffic controllers, who reported difficulties understanding the intermixed voices of many pilots emanating from a single loudspeaker (Kantowitz and Sorkin, 1983). Research has shown that removing spatial dimension from sound reduces intelligibility, memory, and focal assurance (Baldis, 2001). In addition to these difficulties, the loss of spatial cues reduces the ability to leverage the cocktail party effect (Arons, 1992). Thus there are two aspects of human auditory function we can address to improve audio conferencing — “the cocktail party effect” and the enhanced comprehension, memory, and enjoyment resulting from a natural sounding, spatialized, audio environment.

## 2 The Cocktail Party Effect

The *cocktail party effect* is a colloquial term for *selective attention* — the ability to focus one’s listening attention on a single talker in a noisy environment consisting of many conversations. [Arons \(1992\)](#) suggests several ways in which this effect may be viewed. From a listener’s point of view, this ability is nearly effortless. However, from a psychological or physiological perspective it is a very complex interaction of the physical sound waves, the listener’s auditory system, and the human cognitive function. Many aspects of these interactions are poorly understood. The gap in understanding makes it nearly impossible to duplicate this human ability using a computer — resulting in the relatively poor performance of even the most advanced automated speech recognition systems in a noisy environment.

Early work on the cocktail party effect focused on the ability of humans to differentiate between simultaneous speech streams using one or two ears ([Cherry, 1953](#)). Cherry investigated how human listeners can distinguish separate voices in a cacophony of sound. He suggested five factors:

- The voices come from different directions
- Lip-reading, gestures, and the like
- Different speaking voices, mean pitches, mean speeds, male vs. female, etc.
- Different accents
- Transition probabilities (based on subject matter, voice dynamics, syntax, etc.)

A simple setup can eliminate all but the last factor, revealing its importance. Two messages from the same speaker are mixed and played back simultaneously. When experiments were carried out, it was found that listeners could still separate the voice streams. In a variety of experimental configurations, listeners were asked to “shadow” one of the audio streams; the participant repeated the words after hearing them on the recording. The success at this task was evidence that some linguistic properties help listeners maintain focus on a single vocal stream — without this cognitive function, we would “hear” at worst, babel, and at best, a nonsense stream of words chosen from either of the two simultaneous messages.

Additional experiments using two ears were carried out to investigate the ability of listeners to monitor the content of a “rejected” stream heard in one ear, while shadowing a main audio stream heard in the other ear. It was found that subjects were comfortable with the task, and were able to switch focus between streams at will. Simple comprehension questions were asked about the rejected stream after the shadowing was complete. It was found that subjects could not recall items not attended to. In fact, subjects generally did not even notice if the rejected audio stream was played backward, or if the language was switched to German. However, they did notice a change in speaker gender, or a change to a pure tone. This seems contradictory to the existence of the cocktail party effect. [Stifelman \(1994\)](#) explains this contradiction by challenging the experimental setup. The shadowing task is too difficult to allow effective monitoring of additional audio streams, she argues. This is evidenced by Cherry’s comment that subjects spoke in a monotone voice when shadowing, showing a lack of comprehension of what they were saying. In fact, they were unable to successfully answer comprehension questions about the shadowed channel either.

[Stifelman \(1994\)](#) removed the difficult shadowing test from the experiment, and attempted to show the existence of the cocktail party effect. Subjects were asked to actively listen to one speech stream closely, in preparation for a comprehension test. Simultaneously, they were asked to passively monitor a second (and in some cases, a third) stream and indicate when they heard a key word, such as a name. It was found that performance on this task was much better. However, both the ability to comprehend and recall content from the first stream and the accuracy of recognition of the key word in secondary streams decreased as the number of simultaneous streams increased.

[\(Arons, 1992\)](#) lists the following as factors important to *auditory scene analysis* — the human ability to segregate different sounds into *streams* within a noisy environment:

**spatial location** Sounds emanating from different locations are separated, sounds from similar locations are grouped into a stream.

**spatial continuity** Sound sources and listeners do not move too far too fast; spatial discontinuities have been shown to break down streams.

**loudness continuity** While not causing segregation, loudness continuity may corroborate other evidence.

**spectral continuity** Features such as fundamental frequency, harmonics, etc. indicate a sound stream.

**visual channel effects** Visual events may assist perceived locality of sound.

**segregation time constant** Streams are segregated and solidified over a period of at least 4 seconds, and it takes a similar time for the history of a stream to dissipate. This allows continuous segregation under ambiguous conditions.

Our ability to perform auditory scene analysis is certainly a factor in the cocktail party effect (Arons, 1992). As the audio in both the work of Cherry (1953) and Stifelman (1994) lacks the spatial location component, neither experiment truly mimics the conditions of a cocktail party. Further investigation is needed into the relative weighting of each of the aforementioned aspects of auditory scene analysis and how they work synergistically to allow for speech segregation and human monitoring of multiple simultaneous speech channels. Despite this, Arons (1992) uses the factors of auditory scene analysis to make several practical suggestions for enhancing the cocktail party effect in an audio conference system with several simultaneous speakers (channels):

- Provide spatial continuity within channels
- Provide spatial disparity between channels
- Associate visual images with sound streams
- Provide fundamental frequency continuity
- Pitch-shift similar voices away from each other
- Filter streams into different frequency bands
- Do not present too much information simultaneously
- Provide enough time for users to fully separate streams

Baldis (2001) provides evidence that providing static visual images associated with sound streams did not enhance the comprehension and recall of subjects in an audio conference environment, although it did enhance their satisfaction with the experience. It is not yet known if live video would have a greater effect.

In the following sections, we will explore several techniques for spatialization of sound that have been used to enhance the experience and productivity of online audio collaboration, and outline

some specific applications that take advantage of some of the suggestions of [Arons \(1992\)](#) to allow for the cocktail party effect and sound stream segregation.

### 3 Spatial Audio

Research ([Baldis, 2001](#); [Kilgore et al., 2003](#)) has shown improved memory, comprehension, and enjoyment when a listener participates in an audio conference in which the voices of others emanate from different points in space. One current direction of computer supported collaborative work is toward telepresence — the recreation of as much perceptual information as possible in a remote location; making participants feel “virtually there”. As audio is inherently spatial, by reproducing this characteristic, we can enhance the experience of telepresence ([Evans et al., 1997](#)). The placement of different voices in space can be accomplished “within the head” of the listener, using *perceptual synthesis* techniques to modify audio signals delivered using ear phones, or by reproducing the original *sound field* using multiple speakers. Each of these spatialization techniques offers benefits and drawbacks.

#### 3.1 Perceptual Synthesis

There are two main techniques for creation of “within the head” spatial audio: binaural recording and audio manipulation. They differ in their computational requirements and recording technique.

Binaural recordings can be created using a dummy fitted with microphones in the ears. The structure of the human ear is duplicated as much as possible, and the microphones effectively capture any sound localization information present. A human listener then wears headphones and experiences the sound field exactly as does the dummy surrogate. The sound may also be delivered by a pair of loudspeakers using filters to eliminate a confounding factor called *left-right cross talk* ([Cooper and Blauck, 1989](#)). The human listener may be fitted with a head-tracking device, so that their head movements may be duplicated by the dummy, giving a more dynamic experience ([West et al., 1992](#)). A variety of wireless head-tracking device exist, usually consisting of a fixed

source and a sensor attached to the listener. Tracking technologies include magnetic, optical, and ultrasonic, and are usually non-disruptive to the listening experience.

Head tracking is also used in generation of artificial binaural recordings, made by modeling the directionally-dependent features present in sound reaching a listener's ears. Commonly an incoming monaural sound signal is modified using a *Head-Related Transfer Function* (HRTF). HRTFs are based on the shape of the head and outer ear (pinna), and are thus different for each individual, and indeed for each ear (Wightman and Kistler, 1989). A HRTF modifies the frequency spectrum of sound and allows for highly accurate modeling of sound location in three dimensions. Generalized HRTFs based on the functions determined using a precise replica of a human head and ear are often used. Alternatively, the HRTFs of a person determined to be an exceptionally sensitive localizer may be used (Wenzel *et al.*, 1993). The exact HRTF applied to a signal is dependent on the relative position of the listener to the virtual location of the sound source — if the listener moves her head, the HRTFs applied to the signal must adapt to maintain spatialization. This is further complicated by cross-talk elimination if loudspeakers are used (Evans *et al.*, 1997). In perceptual synthesis systems, sufficient computational resources must be available to perform the signal processing and adapt to head movement with as little delay as possible.

Kilgore *et al.* (2003) achieve spatialized in the Vocal Village audio conference application with minimal computational requirements and imperceptible delay by using simplified audio processing functions which account for only the two major binaural cues for localizing sounds in space. Interaural Time Differences (ITDs) and Interaural Intensity Differences (IIDs) are modeled and individual effects such as head and ear shape are ignored. This requires only simple delay and gain (volume) modification, instead of manipulation of the frequency spectrum. The result is convincing modeling of location in a horizontal plane, but no modeling of elevation. In addition, head-tracking is not used, as the system is intended for conferencing between users seated at a computer with gaze fixed on the monitor.

The benefits of perceptual synthesis of spatialized audio include the ability to perform further manipulations like those suggested by (Arons, 1992) (see section 2) to increase the segregation of the

voices of participants in an audio conference, simply by modification of the HRTFs. Additionally, this technique does not require special recording equipment or a facility equipped with large arrays of loudspeakers (Kilgore *et al.*, 2003). A drawback is that the experience of spatialized audio using perceptual synthesis, either over cross-talk canceled loudspeakers or over headphones, is a personal one, based on head-tracking of the listener. The spatialization does not scale well to many listeners; the computations must be repeated for each listener. Thus, in some cases, the use of multi-speaker arrays to perform sound-field synthesis may be preferable, as a recreated sound field can be theoretically experienced by as many listeners as can comfortably fit within the field area (Evans *et al.*, 1997).

### 3.2 Sound Field Simulation

Spatialization of audio may also be achieved using an array of loudspeakers (four or more) to accurately replicate the sound field that would exist in a “real” environment. The reproduced sound field may be sourced by a special array of microphones in a remote location, or may be artificially created, for example by directing different vocal streams to different speakers. The goal in sound field synthesis is to recreate the auditory experience as closely as possible to that of a real environment.

The most common implementation of this is called Ambisonics (Malham and Myatt, 1995). Ambisonics techniques include the cinema technologies THX, Dolby Digital, DTS, and the international standard 5.1 Surround (Rumsey, 2001). A highly sophisticated sound field synthesis product named IOSONO is now available. The IOSONO implementation requires installation of panels with eight speakers each all around the periphery of the listening environment (Brandenburg, 2004). This type of spatial sound setup is more suited to permanent theatrical installation due to its expense.

Ambisonic techniques usually require special recording techniques to capture different channels for each speaker. The number of audio channels (four or more for full 3-D spatialization) increases the bandwidth or amount of storage media needed to transmit or store the sound, as opposed to

the monaural or binaural recordings used to create spatialization in perceptual synthesis. Additionally, even with as few as four speakers, sound field synthesis equipment becomes expensive and is not very portable. In fact, the most common implementations are permanent installations, as opposed to the desktop conferencing using a standard personal computer possible using perceptual synthesis (Evans *et al.*, 1997). The benefit of recreation of a sound field with multiple speakers is that many listeners may simultaneously experience the sound, and computationally intensive and individualized manipulation of the signals is not required.

### 3.3 Comparative Experiments

Evans *et al.* (1997) presents the most comprehensive set of spatialization performance comparators between perceptual and sound field synthesis to date. Two areas of performance are examined — the effectiveness at which sound is spatialized and the quality with which the words being conveyed are understood. Spatialization performance evaluation was carried out using a straightforward perceptual test. Participants were seated and asked to listen to pairs of sentences presented using each of three techniques: ambisonics, very detailed HRTF model, and more computationally simple HRTF model. The ambisonics presentation used a square array of loudspeakers, while the HRTF techniques used 2 cross-talk canceled loudspeakers. Audio was spatialized to come from one of 12 azimuths on the horizontal plane, separated by 30 degrees and corresponding to the hours of a 12-hour clock. Participants were asked to call out the “hour of the clock” from which the speech stimulus appeared to emanate. The results show that for both perceptual synthesis techniques, the apparent azimuth corresponds well to the intended azimuth. The majority of incorrect localizations differ by a single “hour”. In the case of both high quality and lower quality HRTF, errors of front-back reversal were also seen. Front-back reversals were less pronounced for the ambisonic sound field synthesis, but the spread of incorrect localizations was greater.

Comprehension experiments of (Evans *et al.*, 1997) centered around perceived performance, as judged by the listeners, rather than an objective comprehension test. This is the standard technique laid down by the International Telecommunications Union (ITU). The opinions of listeners were

collected for each of the spatialization techniques used in the localization experiment, and for monaural presentation of audio. The *babble resilience test* asked listeners about the effort required to maintain focus and comprehension on a main voice stream under differing levels of peripheral vocal babble. Results showed, after an extensive Analysis of Variance (ANOVA), that significantly less effort was required for simple HRTF compared to monaural in the presence of low and high amounts of vocal babble. The effort was lower still for the very detailed HRTF model and ambisonics.

From these results, Evans concludes that there are not significant performance differences for perceptual and sound field synthesis. Each technique has benefits and drawbacks. For example, perceptual synthesis requires head-tracking and is best experienced using headphones — which isolates the listener from their “real” audio environment. However, the degree of localization is better. Sound field synthesis requires specially designed equipment and is not portable. Yet, it can be experienced by many participants simultaneously and participants report a higher degree of enjoyment when compared to perceptual synthesis. Thus the choice of technique is dependent on the context of use. It is important to note that there is not yet any agreed-upon standard metrics for measuring the performance of spatialized audio. [Evans \*et al.\* \(1997\)](#) suggests a more flexible “range of directions” approach be adopted, which would give an indication of the focus of the localization (*i.e.*, is it heard over a spread of 30 or 60 degrees?), as well as its apparent location.

### 3.4 Memory and Comprehension Experiments

[Logie \(1995\)](#) describes a breakdown of memory into two categories: long term and short term working or working memory. The working memory is further broken down into two distinct systems: the phonological loop and the visuo-spatial sketch pad (VSSP), responsible for temporary retention of visual and spatial information. Experiments have been carried out to test the hypothesis that by adding spatialization to audio conferences, workload is transferred from the phonological loop to the visuo-spatial sketchpad and the apparent capacity of working memory will be increased. A series of experiments designed to measure effects of spatialization on memory, comprehension, and preference have been reported for perceptual synthesis ([Kilgore \*et al.\*, 2003](#)) and sound field

synthesis (Baldis, 2001).

Baldis (2001) designed a series of within subjects experiments to test memory, focal assurance<sup>1</sup>, and preference. Participants were asked to listen to short prerecorded audio conferences with four speakers. The presentation of the audio was varied from non-spatial, co-located spatial, and scaled spatial. Co-located spatial indicated a small separation between speaker location — loudspeakers were placed at 15, 5, -5, and -15 degrees, directly above each of 4 static conferee images. The scaled spatial scenario had the speakers spread out at 60, 20, -20, and -60 degrees. The selection of conference and presentation method was randomized. A training session was carried out to familiarize participants with the voices and names of conferees. After each conference, the participants answered several questionnaires:

- Memory was evaluated using speaker identification: statements from the conference were given and the participant was asked to identify which speaker said it, and their confidence in the answer.
- Focal assurance was evaluated by asking participants to outline the views of each conferee on the topic of discussion, and indicate how confident they were that their assessment was accurate. They were also asked how well they thought they understood what was being said and how difficult it was to determine which conferee was speaking.
- Preference was measured using an end-of-study questionnaire in which participants ranked the listening experience of each conference, and reported on how spatial cues did or did not help their comprehension.

Results showed that for sound field synthesis, participants' performance and confidence on the speaker identification task was higher than for non-spatialized audio, indicating that spatialized audio enhanced memory of the conferences. In addition, focal assurance scores and listener preference were significantly higher for the spatialized audio conditions, compared to the non-spatial. Despite

---

<sup>1</sup>Focal assurance cues include information about conferee participation, such as who is speaking or who is asking questions. Focal assurance is considered a good measure of the existence of a "group space".

the previous results of Egan *et al.* (1954), that intelligibility increases with spatial separation of voices, there was no significant improvement for the speaker confidence or focal assurance scores for the scaled vs. co-located conditions. This may be due to the ease of the task — the signal to noise ratio was high, and comprehension was not very difficult for any of the conditions. Baldis suggests that perhaps under more challenging conditions, the increased spatial separation would help. Finally, participants noted that less effort was required to comprehend the spatial audio, lending support to the theory that when spatial cues are included, work can be “time shared” between the phonological loop and VSSP in working memory.

These experiments were later repeated using a perceptual synthesis system called the Vocal Village (Kilgore *et al.*, 2003). Results showed that the spatialization did not lead to increased memory in this case. However, when participants were given the option to move individual voices around in the virtual space, there was an increase in memory. Participants reported that two of the conferee voices were similar, and by separating them as much as possible, the task was made easier. As in the work of (Baldis, 2001), participants preferred spatialized audio over non-spatial, and their perceived confidence in their focal assurance and memory increased dramatically.

## 4 Spatial Audio Conferencing Applications

Creative and different applications of spatial audio in audio conferencing and online interactive environments have emerged over the past decade. Spatial audio first reached the public in the form of ambisonics for movie theatres, later becoming the brand names THX, Dolby Digital, DTS, IMAX, and the international standard, 5.1 surround (Rumsey, 2001). Later, these technologies migrated to home theatre and then to spatial audio for gaming (O’Neill, 2003). Now that these games are going live online, research is being conducted into providing vocal interaction between avatars in the virtual 3D environment. Simple perceptual synthesis techniques have been applied to provide more a realistic audio environment in which players whose avatars are closer to one another may chat, while hearing other nearby conversations as whispers — an approximation of

the cocktail party effect (Yamazaki and Herder, 2000). Processing and data transfer induces some sound delay, especially when bandwidth is used to transfer additional information, for example about the avatar locations in the virtual environment. Although promising, virtual environments, movies, and gaming are intended for entertainment, and not collaborative work. However, research in these high-profit sectors can be used to the advantage of computer supported collaborative work.

We define spatialized audio conferencing as those collaborative work spaces that include spatial audio and may include a video aspect, but do not have a 3-dimensional virtual reality component. These systems are usually specialized for collaborative work and may provide for shared document workspaces or whiteboards. Spatialized audio conferencing thus includes “desktop conferencing” using a standard desktop computer and perceptual synthesis, and the many emerging forms of collaborative work spaces that use specialized equipment and sound field simulation. Several spatialized audio conferencing applications have been specifically developed for collaborative work between participants possibly remote from one other. Two of the most interesting are the Hydra system (Sellen *et al.*, 1992) and the Vocal Village (Chignell, 2004).

The lines between virtual environments (used mostly for gaming and casual chat) and spatialized audio conferencing are blurring. Take for instance the Vocal Village, an online desktop audio conference system that uses perceptual synthesis to position conferees voices in different locations in space. Similar to virtual environments, the Vocal Village allows for rearrangement of the virtual positions of participants within the audio space. However, the listener in the Vocal Village controls the position of the remote conferees, as opposed to the independent movement of avatars in the virtual environment. Listener control of position has been shown to be beneficial to memory and comprehension (Kilgore *et al.*, 2003).

Enhanced teleconferencing may also allow for parallel and side conversations. The Hydra system is one example. In this system, each conference participant is represented by a separate video screen, camera, microphone, and speaker. These units can be arranged on a desk in a configuration reminiscent of conferees sitting around a table. In addition to spatial audio through sound field synthesis, this system has spatial video, allowing for eye contact, glancing, etc. Each conferee has

a unique view of every other. For example, to have a private side conversation using Hydra, one simply leans closer to the video screen representing the user one wishes to speak to (Buxton *et al.*, 1997). Parallel side conversations are also supported in the Vocal Village and may be initiated between any users. Both systems allow for passive monitoring of the main discussion, thus giving the listener the ability to use the “cocktail party effect”. However, unlike a face-to-face meeting or a virtual environment, the existence side discussions in an enhanced audio conference like the Vocal Village may not be obvious to participants not involved in that side discussion. When standard social constructs, such as the ability to see who is having a side conversation with whom, are not preserved, participants may be wary of the content of the main conversation (Buxton *et al.*, 1997). Further research into the effects of “secret” side conversations on the quality of the main discussion is encouraged.

## 5 Conclusion

We are moving constantly toward a different kind of work environment. Interactions between co-workers are changing as work groups become more and more distributed around the world. However, the quality of the interaction through currently commonplace monaural audio and video conference systems lacks the realism of a face-to-face meeting. Participants cannot initiate side conversations; they cannot whisper to one other while monitoring the main conversation. Indeed, they often cannot even comprehend what is being said as too much vocal information is squeezed into a monaural channel. Through a better understanding of human auditory abilities, there exist opportunities to improve the performance and experience of collaborative work using audio conferencing. These include the spatialization of audio and the ability to use “the cocktail party effect”. Systems are emerging that will provide the benefits of spatial audio, but many of the ways in which our experience may be further improved have not yet been fully investigated. For example, users of the Vocal Village report that comprehension and memory are reduced when two or more voices sound the same (Kilgore *et al.*, 2003). Arons (1992) suggested that we pitch shift similar

voices apart to provide better resolution. Studies into how pitch shifting would affect users have not been reported. For example, would it be confusing if a co-workers voice sounded different in an audio conference? The inclusion of spatial audio, either through perceptual synthesis on our desktop computers or sound field synthesis in specially designed conference rooms is clearly a step toward telepresence, but there are many exciting opportunities for future work.

## References

- Barry Arons. 1992. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, pages 35–50, July. [1](#), [2](#), [3](#), [4](#), [5](#), [6](#), [13](#)
- Jessica Baldis. 2001. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 166–173, Seattle, U.S.A., April. ACM Press. [1](#), [4](#), [5](#), [10](#), [11](#)
- Karlheinz Brandenburg. 2004. Iosono wave-field synthesis system. <http://www.iosono-sound.com>. [7](#)
- William Buxton, Abigail Sellen, and Michael Sheasby. 1997. Interfaces for multiparty videoconferences. In K. Finn, A. Sellen, and S. Wilber, editors, *Video Mediated Communication*, pages 385–400. Erlbaum, Hillside, U.S.A. [13](#)
- E. C. Cherry. 1953. Some experiments on the recognition of speech with one or two ears. *Journal of the Acoustical Society of America*, 22:61–62. [2](#), [4](#)
- Mark Chignell. 2004. The vocal village. <http://www.vocalvillage.net>. [12](#)
- D. H. Cooper and J. L. Blauck. 1989. Prospects for transaural recording. *Journal of the Audio Engineering Society*, 37:3–19. [5](#)
- J. P. Egan, E. C. Carterette, and E. J. Thwing. 1954. Some factors affecting multi-channel listening. *Journal of the Acoustical Society of America*, 26(774–782). [11](#)
- Michael Evans, Anthony Tew, and James Angus. 1997. Spatial audio teleconferencing - which way is better? In *Proceedings of the International Conference on Auditory Display*, Palo Alto, U.S.A., November. [5](#), [6](#), [7](#), [8](#), [9](#)
- S. Handel. 1989. *Listening: An Introduction to the Perception of Auditory Events*. MIT Press, Boston, U.S.A. [1](#)
- B. H. Kantowitz and R. D. Sorkin. 1983. *Human Factors: Understanding People-System Relationships*. John Wilen and Sons. [1](#)
- Ryan Kilgore, Mark Chignell, and Paul Smith. 2003. Spatialized audioconferencing: What are the benefits? In *Proceedings of the 2003 IBM Centre for Advanced Studies Conference on Collaborative Research*, pages 135–144, Toronto, Canada. [5](#), [6](#), [7](#), [9](#), [11](#), [12](#), [13](#)

- R. H. Logie. 1995. *Visuo-spatial Working Memory*. Erlbaum, U.K. 9
- D. G. Malham and A. Myatt. 1995. A 3-d sound spatialization using ambisonic techniques. *Computer Music Journal*, 4(19):58–70. 7
- Cliff O’Neill. 2003. Surrounded by sound. <http://www.gamechronicles.com/features/surrounded/body.htm>. 11
- Francis Rumsey. 2001. *Spatial Audio*. Music Technology Series. Focal Press, Boston. 7, 11
- Abigail Sellen, William Buxton, and John Arnott. 1992. Using spatial cues to improve videoconferencing. In *Proceedings of the 1992 Conference on Computer-Human Interaction (CHI ’92)*, pages 651–652. ACM Press. 12
- Lisa Stifelman. 1994. The cocktail party effect in auditory interfaces: A study of simultaneous presentation. Technical report, MIT Media Laboratory, September. 3, 4
- E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. 1993. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94:111–123. 6
- J. E. West, J. Blauert, and D. J. MacLean. 1992. Teleconferencing system using head-related signals. *Applied Acoustics*, 36:327–333. 5
- F. L. Wightman and D. J. Kistler. 1989. Headphone simulation of free-field listening i: Stimulus synthesis. *Journal of the Acoustical Society of America*, 85(858–867). 6
- Yasuhiro Yamazaki and Jens Herder. 2000. Exploring spatial audio conferencing functionality in multiuser virtual environments. In *Third International Conference on Collaborative Virtual Environments*, pages 207–208, San Francisco, U.S.A., September. ACM Press. 12