AMBISONIC SPATIALIZATION FOR NETWORKED MUSIC PERFORMANCE

Michael Gurevich, Dónal Donohoe, Stéphanie Bertet

Sonic Arts Research Centre
Queen's University Belfast
Belfast BT7 1NN, UK
{m.gurevich,ddonohoe01,s.bertet}@qub.ac.uk

ABSTRACT

In spite of recent widespread interest in network technologies for real-time musical collaboration between distant locations, there has been little focus on spatial audio in such applications. We discuss the potential for dynamic spatialization in the context of network music collaboration, in particular through the use of Higher Order Ambisonics. We describe a platform for real-time encoding, streaming and decoding of spatial audio using Ambisonics, and provide details of two case studies of creative applications built on top of this platform. We demonstrate that Higher Order Ambisonics is a viable and effective means for real-time, simultaneous spatialization in multiple locations, and that it enables a range of creative uses that explore the nature of space, distance and location in networked performance.

1. INTRODUCTION

There has recently been substantial attention on the use of networks to support musical collaboration between performers in remote locations, facilitated primarily by the advent of dedicated research networks that can support the large bandwidth, low-latency packet routing, and guaranteed quality of service (QOS) that such applications require [1, 2, 3]. Among typical configurations for networked musical performances are those in which performers in two or more distant locations attempt to play a fixed composition or some kind of improvisation together in real-time, what Weinberg [4] describes as the "Bridge" approach. In these cases, it is well known that the inevitable latency due to the physical transit time of network packets impacts upon performance [5, 6, 7, 8, 9, 10]. As such, several systems have tried to account for these latencies through design or composition [11, 12, 13].

What these systems have in common however, is that in a given location, the remote performers are conceived as sound sources that are "piped in" to the local space. Although it is not frequently discussed in the literature, it appears that in most cases, the remote streams are played back over loudspeakers with little attention paid to the spatial configurations of the remote or local spaces, or the interactions of the two. In order to minimize the potential for echoes or feedback caused by the remote stream being picked up by local microphones and re-sent back to the source location, performers are typically close-miked, meaning that their spatial representation in the remote space is determined by arbitrary amplitude-based stereo panning. Conceptually, this leads to a difficulty in the representation of the location of remote performers. We conceive of them as elsewhere, but the lack of a specific "source" in the local space means that they are everywhere or nowhere in particular. In spite of the rhetoric of connectedness, there is a strong sense that we are playing with someone who is "elsewhere" – somewhere other than here – as opposed to someone with whom we are virtually sharing the same space, together.

Several studies have considered the consequences of network topologies on conceptualizations of space and location, but these typically avoid the specific question of spatial location within a room. Weinberg's framework for interconnected musical networks [4] considers a variety of network topologies in which "every musical parameter, such as pitch, rhythm, timbre, or dynamics, is a candidate for autonomous as well as interdependent control" but notably does not consider spatial location or impression among these. Föllmer [14] discusses the "problematics" of space and presence – that networks themselves are highly dimensional and confound the notion of physical presence – but fails to acknowledge that at the end of the network there is a real person in a real space who is hearing real sounds. Rebelo [15] further explores the network as a space in itself, citing a series of pieces that explicitly play with the association between data propagating on a network and sound or vibrations propagating in physical space [16, 17, 18, 19]. Along these lines, Rebelo's own piece Netrooms uses the network as a delay line to explore "the juxtaposition of multiple spaces as the acoustic, the social and the personal environment becomes permanently networked" [20].

In spite of the rhetoric of the network as a dynamic and unpredictable virtual place, one of the things that a designer of a networked music performance system *can* control is the displayed spatial location of remote (and local) audio streams – where, in real, spatial terms, should the perceived "source location" of the audio playback be? We propose that by paying closer attention to the spatial aspects of networked music systems, we can not only create an increased sense of togetherness or sharing, but we can also begin to expose and play with the inherent contradictions of space and location in artistic ways.

2. RELATED WORK

Spatiality has been considered among network music transmissions, but it has been discussed most prominently with respect to 1-way streams. In these cases, the most common approach is the reproduction of the spatial impression of a local space in a remote one. Xu et al. 2000 [21] documented a system for real-time transmission of 5.1-channel audio over the Internet, with successful demonstrations from Montreal to New York and Toronto.

The HYDRA system enables synchronous Internet transmission of multi-modal data streams, including multi-channel audio, but there has been little discussion of how these channels are spatialized [22]. An earlier demonstration of HYDRA in 2004 em-

ployed 10.2-channel audio "reproduced live, over 26 speakers" in a 1-way transmission of a string quartet concert [23]. Although the spatial configuration of the 26-channel reproduction is unclear, this performance is notable in its attempt to represent the performers as virtually present in specific locations in the remote venue, at least visually; individual video streams of the members of the quartet were projected, at roughly life-size, onto a stage in the same spatial configuration as the real performers.

Rebelo [15] describes a performance scenario in which video streamed from two remote locations is projected on either side of a central stage of local performers, with audio located accordingly, presumably through amplitude panning. The effect is to give the local audience a sense of two "virtual stages" at either side of the local one, thus conceptually expanding the size of the local space [15]. The same strategy can be simultaneously employed in multiple locations, with the local stage central, although each audience would then experience a different spatial configuration. Rebelo's Netrooms [20], in which audio streams from any number of performers are dynamically mixed and fed back to one another, typically features a single "performance environment" in which an audience can hear the aggregate effect of all the streams. In these performances, the processed stream from each performer is typically assigned to a single loudspeaker, providing each remote participant a virtual point source in the performance environment.

Research in virtual environments, teleoperation and remote presence has acknowledged the utility of spatial audio in facilitating discrimination of remote actors or objects [24, 25, 26, 27]. Such applications tend to be focused on solitary, individual users and therefore most often employ headphone or near-field stereophonic audio displays that are unsuitable for concert audiences or group listening. Other more public cases (e.g. [28]) rely on relatively crude amplitude panning with a small number of speakers. An earlier system (possibly coincidentally) also called *hydra* [29], spatialized participants in a telepresence system by assigning each to an individual video display and loudspeaker.

3. AMBISONICS FOR NETWORK APPLICATIONS

Among 3D audio approaches, Ambisonics is attractive for network applications. Based on spatial sound field decomposition, Ambisonics and Higher Order Ambisonics aim to reproduce an original sound field and its spatial content. It uses an intermediate B-format to reproduce a sound field, allowing a flexible selection of reproduction systems. The sound field is encoded on spherical harmonics creating spatial ambisonic components. The decoding process recreates the encoded sound field to a playback system either locally or remotely located (Figure 1). A classical playback configuration is an evenly distributed loudspeaker layout.

The first-order ambisonic system is composed of four components, W (omnidirectional component), X, Y and Z (three bidirectional components). These ambisonic components represent the first harmonics of an angular sound field decomposition. Higher Order Ambisonics systems include spherical harmonics of higher orders. The encoding and decoding processes are equivalent to those for the first order. The number of ambisonic components increases with the order: $2M + 1$ components in two dimensions and $(M + 1)^2$ in three dimensions, where M is the ambisonic order. Besides the increasing number of ambisonic components that can be problematic (especially when considering a transmission over the Internet), using higher order ambisonics system brings a more accurate spatial information and a wider reproduction area

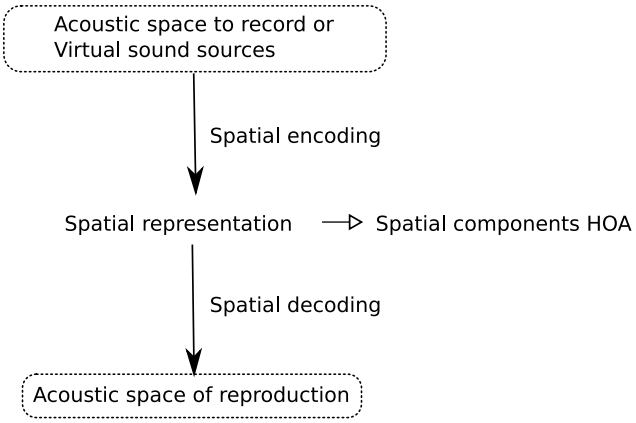

Figure 1: Ambisonics reproduction diagram.

[30, 31]; especially if the reproduction system contains a high number of loudspeakers.

The ambisonic content can be played back in 2 or 3 dimensions depending on the reproduction configuration and encoding process. However the number of loudspeakers depends on the decoding order and should be greater than or equal to the number of ambisonic components. In order to avoid the detent effect (where the sound is pulled toward the closest loudspeaker) Gerzon advises the use of more loudspeakers than the minimum number required [32, 33]. However using a large number of loudspeaker for a given order brings impairment in the reproduced sound field [34, 35]. The optimum number is a trade-off between the number of loudspeaker available and the ambisonic order to decode. The higher number of loudspeakers, the higher ambisonic order can be decoded. The encoding matrix depends on the location of the sources while the decoding matrix depends on the loudspeaker configuration.

Another advantage of having an intermediate format (the decoding process independent of the encoding process) is the possibility to superimpose the signals encoded in different locations and give the same spatial impression decoding the encoded signals to different reproduction systems. For example, at location A a sound source a is encoded in B-format and "positioned" to be reproduced in location B at the defined position. At location B a second sound source b is encoded in B-format and "positioned" to be reproduced in location A at the defined position. The ambisonic components of the two spaces can be "added" together keeping all sound source signals and their spatial information.

Recent studies have investigated transmission of ambisonic sound fields over the network. Noisternig et al. demonstrated streaming of directional instruments over the Internet [36]. The instrument was recorded with an array of microphones to encode it on spherical harmonics. The encoded components were transmitted via the network then decoded to a wave field synthesis rendering system. Hellerud investigated transmission of high audio quality over IP networks using Higher Order Ambisonics reproduction [37]. BBC R&D is exploring Ambisonics technology for broadcasting audio [38], raising a number of technical issues such as the lack of universal decoding system. These studies demonstrate the ability to use Ambisonics for network application, however only 1-way transmissions have been displayed. Our approach combines sound fields encoded in three different spaces at the same time.

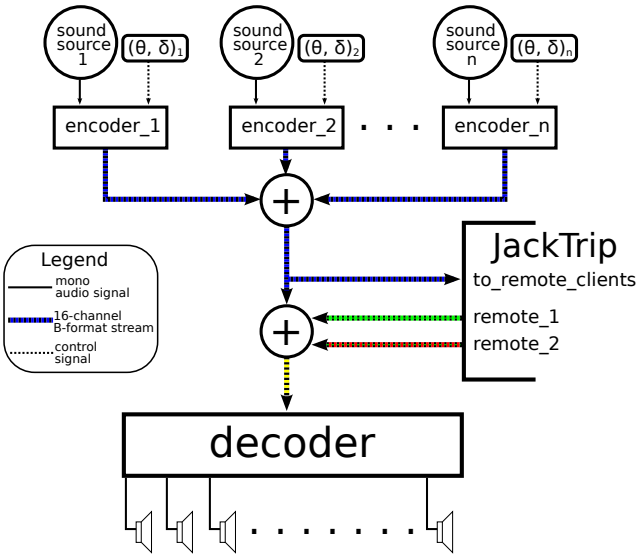

Figure 2: Block diagram of local encoding, decoding and streaming system.

Each space has one specific loudspeaker layout where the encoded sound field is played back using a dedicated decoding matrix.

4. IMPLEMENTATION

The introduction of multichannel ambisonic-encoded signals to a bi-directional network stream, as opposed to the usual mono or stereo signals used in network audio, poses a challenge in terms of processing power and network capabilities. In this section we describe an implementation of a system for streaming 3D 3rd-order ambisonic components between multiple locations.

4.1. Ambisonic Encoding/Decoding

In recent years, practical experiments and developments in Ambisonics have been carried out using a variety of audio processing environments including Csound [39], Pure Data [40], SuperCollider [41] and Max/MSP [42]. Our implementation uses Max/MSP to perform real-time encoding and decoding of the ambisonic components. The encoder is a Max/MSP abstraction running on a local computer that takes as inputs a (local) mono source signal and a pair of azimuth (θ) and elevation (δ) angles representing the sound's desired spatial location. Coefficients of the encoding matrix are determined from θ and δ according to the semi-normalized Furse-Malham formula [43]. The encoder outputs up to 16 B-format component signals which accommodates 3rd-order reproduction in 3D. One instance of the encoder is required for each local sound source, allowing for flexible scaling of the number of local sources according to available processing power.

The 16-channel B-format streams from the local encoders are summed with incoming remote ambisonic streams and routed to a single local decoder also implemented in Max/MSP. The decoder at each location must be configured with the azimuth, elevation and radii of the loudspeakers relative to the center of the room. The decoding matrix contains the pre-determined loudspeaker angle information. Applied to the encoded signals, the matrix outputs

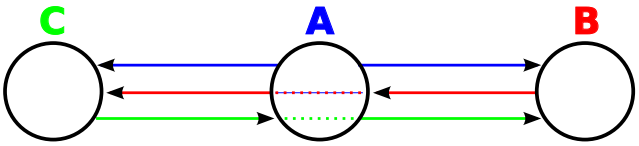

Figure 3: Streaming Configuration. Location A serves 2 separate connections with locations B and C, forwarding their streams to each other.

unique gain-weighted signals for each individual loudspeaker. To achieve spherical unity on the playback surface, where loudspeakers have varying radii from the center, precise signal delays and gain compensation are applied to the local output signals.

4.2. Audio Streaming

In addition to being routed to the local decoder, the 16 B-format component signals from each instance of the encoder are routed to the Jack Audio Connection Kit [44]. Jack is a cross-platform audio host application that enables flexible, near-zero-latency audio connections between applications on the same host computer. Jack allows the spatial audio processing in Max/MSP to interface with JackTrip for network streaming. JackTrip [45] is a high-fidelity audio streaming engine that features a ring-buffer-based, multithreaded architecture and provides both extremely low latency and packet-redundancy, while supporting any number of simultaneous audio channels. JackTrip depends on stable, reliable network connections in order to achieve low latency, as it uses best effort delivery UDP transport and uncompressed audio. In our applications, streaming between large research institutions, JackTrip was extremely reliable, with very few noticeable dropouts, and multiple, stable 32-channel connections for up to 8 hours at a time. A block diagram of the complete encoding, decoding and streaming system for one local computer is shown in Figure 2.

4.3. Issues

Recent improvements to JackTrip that support multi-client servers which can be used as "mixing hubs" [3] are not yet widely available, therefore JackTrip currently supports only peer-to-peer connections. Simultaneous streaming of ambisonic-encoded streams between more than two locations thus requires multiple JackTrip server instances. In our tests with simultaneous streaming between 3 locations, we had most success designating the computer in one location as the "host", which ran 2 instances of the JackTrip server.

If we describe the host as location A, then computers at locations B and C each connect to one of these server instances, sending their 16-channel ambisonic stream to location A. The computer at location A then forwards B's stream to C and C's stream to B, as well as its own stream to both. Locations B and C therefore each receive 32 channels of uncompressed audio from location A, representing 2 separate sets 3rd-order ambisonic components. At each location, all 3 sets (1 local and 2 remote) of 16 components are summed before decoding, requiring only a single instance of the decoder at each site. This arrangement is depicted in Figure 3.

Both ends of a JackTrip connection must share a common sample rate and buffer size which are fixed at Jack startup [3]. Finding an optimum buffer size was challenging in our applications. The Jack buffer size determines the maximum signal vector size in Max/MSP, as well as the internal buffering and packet

length in JackTrip. Longer buffers are computationally more efficient in terms of DSP cycles, but incur longer latencies [45]. Initial experiments with 1st-order Ambisonics (four channels of audio) were successful with 256-sample buffers, but the 32-channel connections necessary for 3-way streaming failed. This was apparently due to the fact that JackTrip wraps the current buffer of all channels in a single UDP packet; 32 channels of 256-sample buffers caused JackTrip to overrun the maximum UDP packet length. The system ultimately worked reliably in a 3-way simultaneous streaming configuration of 16-channel ambisonic streams represented in Figure 3 with 128-sample buffers at 44.1 kHz.

5. CASE STUDIES

We present 2 case studies of creative applications built on top of the Ambisonics-based networked spatial audio system described in Section 4. The first is a game-like interaction that promotes direct interaction between remote users as they manipulate the spatial trajectories of live sounds. The second lies somewhere between a performance and an installation, in which members of the public can contribute their voices to create a shared environment of sounds moving within and between remote locations.

5.1. NetSpace: Collisions

NetSpace: Collisions offers public users a first-hand experience of dynamically controlling the position and movement of a sound in 3D space. However, this space is not exclusive to the user – it is occupied as well by other sounds which have originated elsewhere. The motions of these foreign sounds are controlled by users in remote sites and projected locally in real time; similarly the local sound is observed by the users at distant locations. Effectively, participants virtually share a common spatial acoustic environment. The system is based on the underlying technology discussed in Section 4, in which all sounds are projected through a 3D array of loudspeakers using Higher Order Ambisonics. The reliability and low-latency streaming capabilities of JackTrip made the experience truly interactive rather than just participatory. As a result, as users discover that the system responds when their sounds approach and intersect in space with those from the distant locations, they can play with deliberately creating or avoiding "collisions".

5.1.1. System Overview

In *NetSpace: Collisions* sounds are generated from a microphone that samples and loops the user's voice. Motion of the sampled sound is controlled via an external controller in the center of the space, or alternatively a screen-based graphical interface, which also serves as a visualization. On September 8, 2010 the system was exhibited in a three-way connection between the Sonic Lab at SARC, Queen's University Belfast, the Listening Room at CCRMA, Stanford University, and Studio 4 at IRCAM, Paris. Although the system is conceived to employ a physical controller at each site, due to logistical problems this exhibition featured the purpose-built light-sensitive dome controller described below at one location, and the GUI interface only at the other two.

The local user remains in the centre of the reproduction space where they have full dynamic control over the spatialization of their input sound. Being located within the sweet spot allows them to enjoy the benefits of precision angular discrimination of Higher Order Ambisonics. Signal input is sent from the microphone to

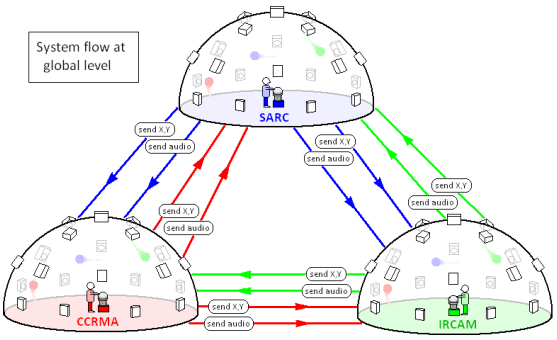

Figure 4: Signal and data flow in *NetSpace: Collisions*

Max/MSP to be encoded, as described in Section 4.1. The local user can control the motion of their sound by dragging a coloured dot representing their source around a GUI depicting the hemispherical playback surface, or by moving a light pen around the hemispherical surface of the controller. The GUI also displays the location and motion of the remote users' sounds. A selector switch offers 1st, 2nd or 3rd-order ambisonic reproduction in order to accommodate playback environments with varying numbers of loudspeakers.

The implementation of user interactivity brings another dimension to *NetSpace: Collisions*, in which the proximity of the different users' sounds to one another affects signal processing applied to their sampled voices. As the ambisonic streams sent between peers contain only encoded audio signals, the spatial coordinate data are sent between sites as well using Open Sound Control (OSC) [46] for display on the GUI and to introduce audio effects based on sounds' relative locations. At each site, a "proximity" function measures the linear distance to the nearest source, controlling the sampling window size and pitch transposition of the input signal. A sound source of decreasing sample duration and increasing pitch indicates "nearby activity" to the user. If sources get too close in space, a collision is registered, signified by a crashing sound at that point in space. Collisions erase the sample memory, requiring users to record a new sample in order to re-establish their interaction with the environment.

5.1.2. Controller Design

The controller, shown in Figure 5, consists of a hemispherical surface, designed around the concept of scaling down the virtual playback surface to a physical model. A point anywhere on surface of the controller relates to the same position on the ambisonic reproduction hemisphere around the user. Moving the control stylus around the surface results in a corresponding movement of the sound source in real-time. The underlying technology is inspired by an interactive spherical display designed by Microsoft [47]. Encased in the tubular base of the controller is a USB webcam, facing directly upward towards the inner surface of the hemisphere. Coupled to this is a door-peephole which widens the viewing angle of the webcam. It is positioned so that its 120 ° field of vision aligns precisely with the hemispheres equator. The hemisphere itself is made of light diffusing plastic, but by concentrating light close to

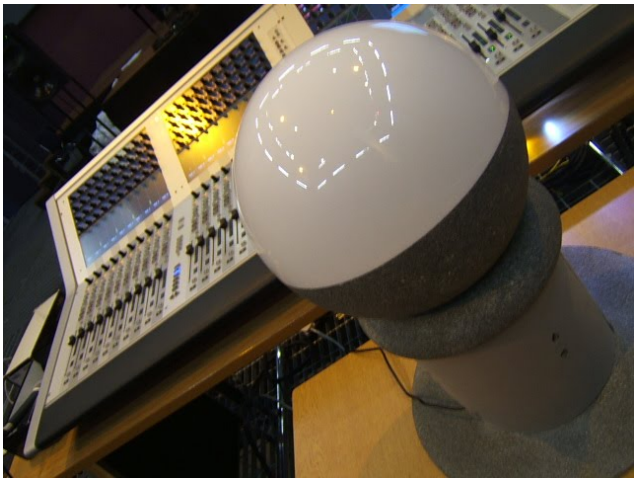

Figure 5: Hemispherical controller in *NetSpace: Collisions*

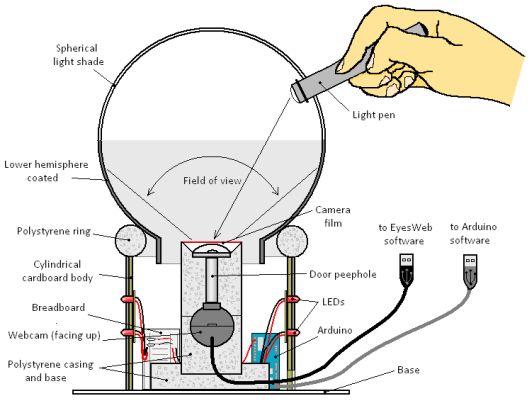

Figure 6: Controller design for *NetSpace: Collisions*

the outer surface, a spot can be seen from the inside and registered on the webcam. For the light source we use a small light-pen which acts as the stylus for dragging the sound around the room.

From the webcam's point of view, a spot of light is seen moving around on a circular 2D plane. To heighten the contrast between the spot and the rest of the surface, exposed film strips were attached to the lens of the peephole. Tracking of the spot's position on the plane is done using the EyesWeb data analysis and processing software [48]. EyesWeb determines the barycenter of a blob that appears in the image from the light pen and outputs its x and y coordinates. The coordinates are then sent as OSC messages to Max/MSP, where they are converted to (θ, δ) ambisonic angles that are input to the ambisonic encoder.

Contributing to the interactive element of the system, the controller also visually informs the user when they have collided with another sound source. When a collision is registered, an array of LEDs at the base of the hemisphere flash rapidly, controlled from Max/MSP via an Arduino microcontroller board [49]. A schematic overview of the controller's design is shown in Figure 6.

5.1.3. GUI Design

The GUI was designed as an alternative control interface which also serves as a visualization to the external controller. It is based on the metaphor of "paint on a canvas," whereby the hemispherical playback surface denotes the canvas and moving sound sources are strokes of paint. The hemispherical surface is displayed in a plan and front elevation view. Using a computer mouse, the user can drag the paintbrush across the canvas, creating a trail of colour that disappears over time. The movement of the sound source follows the path of the brush in real-time. The GUI features an "auto-pilot" algorithm which causes the source to follow a continuous path of random speed and motion, as an alternative to human control.

The GUI also serves as a visualization of the movement of the light pen on the external controller, and simultaneously displays incoming spatial coordinates from remote locations in different colours. The GUI registers collisions that occur on the canvas, by displaying a flashing mixture of colours on the point of contact. Users that have collided are removed from the canvas temporarily until they re-record a new sample.

5.1.4. Discussion

NetSpace: Collisions creates a game-like environment for users in distant locations. Although there is no "objective" to the game – there is no penalty or reward for collisions – we found that participants played with tactics like leading, following and evasion. The microphone provided not only a dynamic, live sonic element, but a means for an additional channel of communication (albeit an imperfect one due to the signal processing) among participants.

The controller's and GUI's direct representations of the ambisonic soundfield allowed users to think of the movements of their sounds directly in spatial terms; there was no conceptual mapping required from control to auditory display. Ambisonics also allowed participants to take for granted that the spatial audio display was the same in their own location as in the others; they did not have to wonder how their manipulations were displayed or perceived in the remote locations. Significantly, there was little attention toward individual loudspeakers; there was a sense that the sounds were located somewhere in space, rather than in one loudspeaker or another. However, participants were aware of the distance between them and there was no pretense of being in the same room. There was a strong sense of shared experience, but one that was located in the interface rather than in the space itself. The result was an immersive, shared experience in which participants felt like they were directly collaborating and competing with one another; they were aware of the commonality of spatial and sonic display, but also of the distance separating them.

5.2. Whispering Places

A second piece built on our networked ambisonic platform, *Whispering Places*, is somewhere between a sound art installation and a participatory durational performance. The overall idea is that visitors in different rooms in distant cities can whisper messages into microphones, and these messages will be captured and looped as they move toward the remote locations. All visitors simultaneously hear all the messages from all spaces moving according to the geographical direction of travel between the locations. The piece contrasts the near-instantaneous transport of network packets representing the sounds with the much slower time scale in which humans can perceive sounds moving in space. *Whispering Places* was premiered on September 16, 2010 between the Sonic Lab at SARC, Queen's University Belfast, the Listening Room at

CCRMA, Stanford University, and Studio 4 at IRCAM, Paris.

In *Whispering Places*, there are 4 microphones in each of the 3 spaces, positioned midway along each wall of the space, low to the floor and facing the center of the room such that participants would be facing the perimeter of the room. Cushions and blankets invite visitors to sit or lie on the floor in front of each microphone. A single computer in each location processes the signals of the 4 local microphones. Microphone inputs are buffered and segmented using time-domain analysis of the overall signal envelope. Segments longer than a minimum duration are analyzed for noisiness and spectral tilt in order to separate whispers from voiced speech. If voiced speech is detected, the segment is looped at a quasi-random rate and spatialized along a trajectory until it reaches a target.

An LED mounted on each microphone provides feedback for visitors. When no sound is detected, the LED is off. It turns blue while sound is detected and being recorded to the buffer, then flashes green after the segment ends to indicate that a whisper was detected, or flashes red if a whisper was not detected. The detection threshold was set such that the system erred on the side of missed detections, rather than false positives. It was seen as important in order to preserve the aesthetic integrity of the piece to avoid any possibility of voiced speech or singing from entering the soundscape. The LED feedback provided a way for visitors to learn to whisper in a way that would be reliably detected.

5.2.1. Spatialization

The spatialization in *Whispering Places* is driven by a gravity simulation, implemented in Max/MSP using the pmpd library [50]. The simulation runs on each local machine, controlling the ambisonic encoding of the local microphone signals. In the simulation, a large mass exerting gravitational force is located at each remote location, determined by its GPS coordinates. When a whispered message is detected, the simulation represents it as a small mass in the location of the microphone from which the signal originated, and gives it a "shove" with an initial random velocity, azimuth and elevation. As the message loops, it initially follows this random spatial trajectory, but is also subject to the forces of the gravity wells, which eventually draw the sound toward one or the other remote locations. Damping counteracts the momentum of the initial "shove", ensuring that the sounds will be captured by one or the other gravity wells. When the mass eventually collides with the gravitational source, the message ceases to loop.

The gravitational masses are represented at the same elevation, but the initial random trajectory, which includes an elevation component, ensures that sounds do not only travel in the horizontal plane. As the loudspeaker configuration in one site was hemispherical (there are no loudspeakers below the floor), sounds are bounded by 0° elevation. To simulate sounds moving toward and away from the listening area, a radius parameter is added to the encoder described in Section 4, meaning that sounds are no longer confined to the unit hemisphere. Prior to encoding, sound sources are attenuated by $1/r^2$ and a lowpass filter with linearly decreasing cutoff frequency is applied. Obviously, sounds would rapidly become inaudible if distances between locations were treated literally, so the radii are scaled down by a uniform but arbitrary amount such that messages are perceptually attenuated as they recede toward the remote locations but remain audible.

The aggregate effect is that sounds from Belfast would, after the initial random shove, travel either roughly south toward Paris, or southwest toward California. Due to the relative proximity of Belfast and Paris, sounds originating from California would travel along roughly similar northeasterly trajectories.

At each location, spatialized radio static is played in addition to the messages. The static was pre-recorded by continuously varying the tuning of an analog FM radio over a period of 10 minutes. In the piece, short snippets of this radio static are periodically captured and encoded it at an azimuth, elevation and distance determined by a random process, and streamed to all locations.

5.2.2. System Configuration

A local computer at each site handles the whisper detection and processing for the local microphones. It also runs the spatialization simulation that outputs azimuth, elevation and distance for each looping message at a rate of 20Hz. These parameters are fed to an instance of the 3rd-order ambisonic encoder. For each message, one copy of the 16-channel encoded stream is then sent to the local decoder for local playback. The 16-channel streams for all simultaneously looping messages are summed, creating a single 16-channel output stream that is sent to jacktrip to be distributed to the remote locations. Using a 2 x 3GHz Quad Core Intel Xeon-powered Mac Pro, 5 instances of the encoder per microphone could be run comfortably, for a total of 20 possible simultaneous looping messages per site.

The system employed the streaming topology depicted in Figure 3, with the computer at SARC acting as the "host", running two instances of the JackTrip server and forwarding CCRMA's stream to IRCAM and vice versa. The connections remained stable for approximately 10 hours throughout setup and performance.

5.2.3. Discussion

With *Whispering Places*, we sought to highlight the paradoxes of space, location and direction raised by real-time networked audio interaction, and exploit them in an artistic manner. Visitors reported a sense of immersion and envelopment, and experienced pleasant contradictions in perceptions of presence and location. Some visitors wandered in unaware of the nature of the interaction, not having read the posted description outside the door, and reported an engaging experience arising from not knowing the source or identity of the distant voices.

Indeed, the overall design of the system led to difficulty in determining whether the messages originated from someone in the same room or from far away. In this way, the piece was successful in highlighting the paradox of virtually sharing a real space. The requirement of whispering added a further air of mystery to the experience as whispers lack much of the spectral information we use to identify individual voices. There was a sense not just of wondering *where* these unseen people were, but also *who* they were. The ambiguity of location was aided by dim lighting and the microphone setup – microphones facing the walls helped isolate visitors from one another – but also by the fact that all sounds were spatialized throughout the entire space. Rather than, say, designating one side of the room as Belfast, the other side as Paris and the center as Stanford, we superimposed the spaces on top of one another. Sounds from each location followed separate trajectories as they travelled "toward" the others, but the spatial experience was also shared by everyone simultaneously. Ambisonics allowed us to implement this flexibly due to the independence of the encoded spatial location from the playback environment.

When there were many simultaneous visitors generating a large number of messages, the number of voices made it difficult

to track the sounds as trajectories, although the motion gave them a sense of life and activity. In future versions, a subtle and abstract visualization, perhaps using multiple projection screens throughout the spaces, might help visitors to perceive trajectories of individual messages, which would reinforce the sense of remoteness of the distance between locations.

6. CONCLUSIONS AND FUTURE WORK

We have demonstrated that Higher Order Ambisonics is a feasible and effective solution for implementing 3D spatial audio for networked music applications. It has a significant benefit of allowing for flexible numbers and arrangements of loudspeakers in different sites. Through ambisonic encoding, a source audio signal's spatial location relative to the center of the room can be specified at its originating site, and the ambisonic components can be decoded to reproduce the sound in the same relative spatial location at a remote site. Far from being confined to static spatial locations, these sounds can follow dynamic trajectories specified at their source by control-rate azimuth, elevation and distance parameters.

Ambisonics is a very effective solution in this regard because rooms will never have similar loudspeaker setups and yet it allows the spatial location to be specified at the source. We can conceive of many applications where this capability is desirable, such as representing the spatial locations of close-miked performers distributed throughout a room. One could imagine a networked performance in which the sounds of "ghost" performers from a remote site are interspersed with live performers in the local space. We described 2 applications that go beyond the typical network performance scenario to exploit for artistic purposes the ability to represent multiple, dynamically moving sound sources in 3D simultaneously in different rooms.

A number of practical matters can be improved as our implementation is developed further, such as an enhanced simulation of distance to better represent sounds moving through and away from the space. As an increasing number of playback rooms, including 2 of those used in our study, now support fully 3D loudspeaker locations, a more elegant handling of compatibility between hemispherical and spherical loudspeaker arrangements would be desirable; our system currently does not make use of loudspeakers in the lower hemisphere. The challenges of mutual, multi-site streaming with JackTrip that are currently being addressed [3] will greatly simplify the networking issues we faced.

Beyond these however, we can also begin to explore further ways of exploiting spatial aspects of sound in networked applications. Simultaneously reproducing sound in the same spatial location within rooms in distant sites is but one option; this work is a step in the direction of extended, hybridized and augmented spaces supported by real-time, networked spatial audio.

7. ACKNOWLEDGMENT

We thank Thibaut Carpentier and Jason Sadural, as well as the technical staff at SARC, IRCAM and CCRMA for their assistance in facilitating our work.

8. REFERENCES

[1] C. Chafe, S. Wilson, R. Leistikow, D. Chisholm, and G. Scavone, "A simplified approach to high quality music and sound over IP," in *COST-G6 Conference on Digital Audio Effects*, 2000, pp. 159–164.

[2] C. Chafe, "Tapping into the internet as an acoustical/musical medium," *Contemporary Music Review*, vol. 28, no. 4, pp. 413–420, 2009.

[3] J.-P. Cáceress and C. Chafe, "JackTrip/SoundWIRE meets server farm," *Computer Music Journal*, vol. 34, no. 3, pp. 29–34, 2010.

[4] G. Weinberg, "Interconnected musical networks: Toward a theoretical framework," *Computer Music Journal*, vol. 29, no. 2, pp. 23–39, 2005.

[5] C. Bartlette, D. Headlam, M. Bocko, and G. Velikic, "Effect of network latency on interactive musical performance," *Music Perception*, vol. 24, no. 1, pp. 49–62, 2006.

[6] N. Bouillot and J. R. Cooperstock, "Challenges and performance of High-Fidelity audio streaming for interactive performances," in *Proceedings of the 9th International Conference on New Interfaces for Musical Expression*, 2009, pp. 135–140.

[7] C. Chafe, J. P. Cáceres, and M. Gurevich, "Effect of temporal separation on synchronization in rhythmic performance," *Perception*, vol. 39, no. 7, pp. 982–992, 2010.

[8] X. Gu, M. Dick, Z. Kurtisi, U. Noyer, and L. Wolf, "Network-centric music performance: Practice and experiments," *IEEE Communications Magazine*, vol. 43, no. 6, pp. 86–93, 2005.

[9] A. Kapur, G. Wang, P. Davidson, and P. R. Cook, "Interactive network performance: A dream worth dreaming?" *Organised Sound*, vol. 10, no. 3, pp. 209–219, 2005.

[10] J. Lazzaro and J. Wawrzynek, "A case for network musical performance," in *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2001, pp. 157–166.

[11] N. Bouillot, "nJam user experiments: enabling remote musical interaction from milliseconds to seconds," in *Proceedings of the 7th International Conference on New Interfaces for Musical Expression*, 2007, pp. 142–147.

[12] J.-P. Cáceres, R. Hamilton, D. Iyer, C. Chafe, and G. Wang, "To the edge with China: Explorations in network performance," in *ARTECH 2008: Proceedings of the 4th International Conference on Digital Arts*, 2008.

[13] M. Gurevich, "JamSpace: a networked real-time collaborative music environment," in *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, 2006, pp. 821–826.

[14] G. Föllmer, "Electronic, aesthetic and social factors in net music," *Organised Sound*, vol. 10, no. 3, pp. 185–192, 2005.

[15] P. Rebelo, "Dramaturgy in the network," *Contemporary Music Review*, vol. 28, no. 4, pp. 387–393, 2009.

[16] J.-P. Cáceres and A. B. Renaud, "Playing the network: the use of time delays as musical devices," in *Proceedings of International Computer Music Conference*, 2008, pp. 244–250.

[17] C. Chafe, S. Wilson, and D. Walling, "Physical model synthesis with application to internet acoustics," in *Proc. 2002 Intl. Conference on Acoustics, Speech and Signal Processing*, 2002, pp. 4056–4059.

[18] C. Chafe, "Distributed internet reverberation for audio collaboration," in *Proceeding of the 24th International Conference of the Audio Engineering Society*, 2003.

[19] A. Tanaka and B. Bongers, "Global string: A musical instrument for hybrid space," in *Proceedings: Cast01//Living in Mixed Realities*, 2001, pp. 177–181.

[20] P. Rebelo, "Netrooms," 2010. [Online]. Available: http://netrooms.wordpress.com/

[21] A. Xu, W. Woszczyk, Z. Settel, B. Pennycook, R. Rowe, P. Galanter, J. Bary, G. Martin, J. Corey, and J. R. Cooperstock, "Real-time streaming of multichannel audio data over internet," *Journal of the Audio Engineering Society*, vol. 48, no. 7-8, pp. 627–641, 2000.

[22] R. Zimmermann, E. Chew, S. A. Ay, and M. Pawar, "Distributed musical performances: Architecture and stream management," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 4, no. 2, pp. 1–23, 2008.

[23] T. Rosen, "Is it live or is it Internet2?: Miro quartet shows how technology may change the future of live performances," Oct. 2004. [Online]. Available: http://www.utexas.edu/features/archive/2004/internet.html

[24] S. Aoki, M. Cohen, and N. Koizumi, "Design and control of shared conferencing environments for audio telecommunication using individually measured HRTFs," *Presence*, vol. 3, no. 1, pp. 60–72, 1994.

[25] N. Durlach, "Auditory localization in teleoperator and virtual environment systems: ideas, issues, and problems," *Perception*, vol. 20, no. 4, pp. 543–554, 1991.

[26] N. I. Durlach, B. G. Shinn-Cunningham, and R. M. Held, "Supernormal auditory localization," *Presence*, vol. 2, no. 2, pp. 89–103, 1993.

[27] E. M. Wenzel, F. L. Wightman, and D. J. Kistler, "Localization with non-individualized virtual acoustic display cues," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1991, pp. 351–359.

[28] N. P. Jouppi and M. J. Pan, "Mutually-immersive audio telepresence," in *Proceedings of the 113th Audio Engineering Society Convention*, 2002.

[29] W. Buxton, "Telepresence: integrating shared task and person spaces," in *Proceedings of Graphics Interface '92*, 1992, pp. 123–129.

[30] J. Daniel, "Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia," Ph.D. dissertation, Université Pierre et Marie Curie (Paris VI), France, 2000.

[31] S. Bertet, J. Daniel, E. Parizet, L. Gros, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order ambisonic sound fields : a subjective evaluation involving virtual and real 3D microphones," in *Audio Engineering Society 30th International Conference*, Saariselkä, Finland, 2007.

[32] M. A. Gerzon, "Criteria for evaluating surround-sound systems," *Journal Audio Engineering Society*, vol. 25, no. 6, pp. 400–408, June 1977.

[33] ——, "General metatheory of auditory localisation," in *Audio Engineering Society 92nd Convention*, 1992.

[34] A. Solvang, "Spectral impairment for two-dimensional higher order ambisonics," *Journal Audio Engineering Society*, vol. 56, no. 4, pp. 267–279, April 2008.

[35] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Influence of microphone and loudspeaker setup on perceived higher order ambisonics reproduced sound field," in *Ambisonics Symposium*, 2009.

[36] R. D. Markus Noisternig, Franz Zotter and W. Ritsch, "Streaming directional instruments over the internet," in *2nd International Symposium on Ambisonics and Spherical Acoustics*, May 2010.

[37] E. Hellerud, "Transmission of high quality audio over ip networks," Ph.D. dissertation, Norwegian University of Science and Technology. Faculty of Informatics, Mathematics, and Electrotechnics, 2009.

[38] C. Baume and A. Churnside, "Upping the auntie: A broadcaster's take on ambisonics," in *Audio Engineering Society Convention 128*, 2010. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=15336

[39] D. G. Malham and A. Myatt, "3-D sound spatialization using ambisonic techniques," *Computer Music Journal*, vol. 19, no. 4, pp. 58–70, 1995.

[40] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich, "A 3D ambisonic based binaural sound reproduction system," in *24th International AES Conference: Multichannel Audio*, 2003.

[41] M. Rumori, "Girafe – a versatile ambisonics and binaural system," in *Proceedings of Ambisonics Symposium*, 2009.

[42] J. C. Schacher and P. Kocher, "Ambisonics spatialization tools for max/msp," in *Proceedings of International Computer Music Conference*, 2006.

[43] D. Malham, "Higher order ambisonic systems," Mphil thesis, University of York, England, 2003.

[44] P. Davis, "Jack: Connecting a world of audio," 2011. [Online]. Available: http://jackaudio.org/

[45] J.-P. Cáceres and C. Chafe, "JackTrip: under the hood of an engine for network audio," in *Proceedings of International Computer Music Conference*, 2009, pp. 509–512.

[46] M. Wright, "Open sound control: an enabling technology for musical networking," *Organised Sound*, vol. 10, no. 3, pp. 193–200, 2005.

[47] H. Benko, A. D. Wilson, and R. Balakrishnan, "Sphere: multi-touch interactions on a spherical display," in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*, 2008, pp. 77–86.

[48] A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe, "EyesWeb: toward gesture and affect recognition in interactive dance and music systems," *Computer Music Journal*, vol. 24, no. 1, pp. 57–69, 2000.

[49] "Arduino," 2011. [Online]. Available: http://www.arduino.cc/

[50] C. Henry, "PMPD: physical modelling for pure data," in *Proceedings of the International Computer Music Conference (ICMC'04)*, 2004, pp. 37–41.