INTERCHANGE FORMATS FOR SPATIAL AUDIO

Stephen Travis Pope
FASTLab Inc. & HeavenEverywhere Media
Santa Barbara, California USA
stp@HeavenEverywhere.com

ABSTRACT

Space has been a central parameter in electroacoustic music composition and performance since its origins. Nevertheless, the design of a standardized interchange format for spatial audio performances is a complex task that poses a diverse set of constraints and problems. This position paper attempts to describe the current state of the art in terms of what can be called "easy" today, and what areas pose as-yet unsolved technical or theoretical problems. The paper ends with a set of comments on the process of developing a widely useable spatial sound interchange format.

1. INTRODUCTION

The first known electrically mediated spatial audio performance used Clement Adler's 80-channel spatial telephone and took place at the international exhibition in Paris in 1881. In the 20th century, the development of multi-channel sound recording/reproduction techniques and equipment was carried out in parallel with the evolution of electroacoustic music composition and synthesis hardware/software. It is interesting to note that John Chowning's seminal paper on the simulation of moving sound sources [1] predates his more-widely-cited FM synthesis paper. [2] - [4] also present early results beyond which we have progressed very little in theory, and arguably even in practice.

It is quite unfortunate that there have been only very weak and short-lived standards for state-of-the-art multichannel audio transmission or performance that went beyond the then-current state-of-the-market standards. (Most of the computer music in my collection is on stereo CDs, with a few newer pieces on 5.1-channel DVDs.) Furthermore, there are no standards for even the simplest of extensions, such as the channel-to-speaker assignment for the playback of 4-channel quadrophonic material.

The current state-of-the-market distribution formats are either high-sample-rate, high-resolution (96 KHz, 24-bit) stereo audio (e.g., DVD/A format or www.MusicGiants.com downloads), or CD-format 5.1-channel surround sound. Current DVD playback hardware does not even have adequate bandwidth to play back uncompressed high-resolution surround sound data.

In order to foster the distribution of higher-spatial-fidelity audio, the community naturally seeks to establish a standardized interchange format for 3D- spatial audio scenes. Whether (and to what extent) this format will also be appropriate for use in real-time distributed systems, and how it should provide links for real-time interaction and multimedia synchronization remain open (and interesting) questions.

2. THE CURRENT STATE

A system that allows the interchange of spatial audio between different hardware systems would necessarily consist of (at least) a content editor for creating spatial audio performances, the interchange format itself (file, metadata or stream), and one or more renderers for the chosen spatialization techniques and many-channel output playback/recording formats.

There are three spatial sound rendering techniques that have found wide application in electroacoustic music: (1) vector-based amplitude panning (VBAP), (2) Ambisonics, and (3) wavefield synthesis (WFS) (see [5] for complete references). Each of these make different assumptions about the nature of the sound sources, the existence and impact of a simulated "room" and the features of the listener.

The plentiful software systems for spatial audio performance (see the recent ICMC Proceedings, or those of the special-topic AES workshops), generally adopt exactly one of these processing techniques, and implement it in their output stage (the spatial sound renderer). The two systems reported in the literature support more than one renderer by providing a renderer-independent intermediate layer are the SoundScapeRenderer (SSR) [6] developed at Deutsche Telkom Labs is one, and the CSL system [7] developed by our group at UCSB. Jorge Castellanos, Doug McCoy, Lance Putnam, Graham Wakefield, and Will Wollcott all contributed to this system and used it in their thesis projects. CSL currently supports all three of the rendering techniques listed above, as well as binaural playback over headphones using HRTF-derived filters, and we have evaluated its performance and scalability in a variety of distributed systems, including the UCSB AlloSphere [5].

3. WHAT'S EASY?

The history of spatial audio production includes a number of systems that simulate **point sound sources** with 2-D or 3-D positions, which can be represented as Cartesian (x, y, z) or polar (radius, azimuth, elevation) positions relative to a stationary origin (often chosen to be the middle of the ideal listener's head).

Moving sound sources can be simulated using a variety of techniques, and the motion itself can be described using source trajectories represented as time-stamped geometry updates. The renderer can optionally perform **source position interpolation**. Graphical editors for source positions have been built going back to the late 1970s [1] - [3].

Using the client/server model, a **distributed system** can easily be built that **streams sampled audio** (e.g., using RTP or SDIF as the interchange protocol) in parallel with **control-rate geometry updates** (e.g., using OSC) from synthesis and interaction clients to a spatial

sound rendering server. Given such a system, one could simply store the audio and control streams along with the time-stamp information necessary to correlate them (possibly along with other metadata), and call this combination an interchange format for spatial audio scenes.

More complex renderers might choose to model sources with orientations as well, enabling directional, frequency-dependent **source radiation patterns**.

It is also a solved problem to compute **early reflections** based on a desired room model stored as a set of surfaces.

4. WHAT'S HARD?

The bulk of the recent research literature has concentrated on technical problems that are specific to each of the three standard rendering techniques, rather than on higher-level and cross-domain issues. Aside from these issues, and from those that arise in building renderer-agnostic middle-stage processing, there are still several areas that pose problems in this area.

The modeling and processing of **diffuse and distributed sound sources** is largely unaddressed, save in systems that allow for unspatialized "global" sources.

The **visualization** of, and **interaction** with, true **dynamic spatial audio scenes** (including, of course, dynamic "rooms" within these scenes) is still difficult.

Spatial audio playback over **low-channel-order systems** (e.g., stereo loudspeakers or headphones) is an active areas of research, with techniques for both output formats being developed. This is more a renderer issue than an interchange format issue.

Lastly, Developers in several application areas have had to confront the **scalability issues** that arise when one wants to support many sources, many output channels, or rapid source movement. The performance can be characterized in terms of compute-load, network bandwidth, and distributability (coupling between servers) [5]. The common rendering techniques have quite different performance degradation behavior, though this is also a renderer implementation issue rather than an interchange format issue. We can, for example, generalize about the scaling of VBAP vs. Ambisonics for the support of many output channels, since VBAP would require a dynamic network in which sources are switched between servers each processing a geometrical region of the output space, whereas with Ambisonics the same Nth-order channel buffers are streamed from the encoders to the decoders.

5. WHAT'S TO DO?

Given the fact that I included "an interchange format for spatial audio scenes" in the section on "what's easy" above, the question arises of why we are even discussing it any more. The answer is that there are a number of unanswered questions—some related to the models that should underly said interchange format, and some related to the storage and processing of data store using the format.

The models that are required include minimally:

monophonic sound source with position, orientation, and directional and frequency-dependent radiation pattern; and

room with wall characteristics and listener position.

The first model would include enough metadata to allow the renderer to locate or connect to the (live or stored) sound source sample stream, and to position and orient the source in its virtual space. The second model allows the renderer to compute the simple processing parameters (using the listener position data), and (optionally) to configure one or more reverberators based on the simulated room and source/listener geometry.

In the degenerate case, this could all be stored in an extended-format AIFF, SDIF, OSC, or XML file. (We already have the data structures in CSL [6]). There are several concrete proposals for this in the references of the other panel papers.

We should develop a survey the systems that use related formats (WONDER, HRTF-players, SpatDIF, MPEG-4, X3D, OpenAL, VRML, etc.) to refine our base set of requirements and feature set ideas.

The challenges come when one tries to offer true renderer-independence while supporting the features that are natural for each of the well-known rendering techniques (e.g., plane wave sources on WFS or low-order sources in mixed-order Ambisonics). Nevertheless, any of the pre-existing proposals could serve as the basis of a request-for-comment aimed at establishing a more widely used standard interchange format for spatial audio content.

6. CONCLUSIONS

It will be obvious to the reader that this paper takes a very pragmatic and low-level approach, and that this must be merged with the higher-level perspective based on auditory spatial schemata as is offered by the other panel participants. In this respect, the current contribution is more in line with the renderer-agnostic declarative 3D audio formats that have arisen in the game development and ICAD communities. Our own work at UCSB is centered on the provision of a flexible distributed interactive software framework for a very-large-scale spatial audio rendering system in the UCSB AlloSphere space.

7. REFERENCES

[1] Chowning, J. M. "The Simulation of Moving Sound Sources". *JAES* 19(1): 2-6; 1971. reprinted in *CMJ*, 1(3) 48-52, 1977.

[2] Moore, F. R. "A General Model for Spatial Processing of Sounds." *CMJ*, 7(6): 6-15, 1983.

[3] Moorer, J. A. "About this Reverberation Business." *CMJ*, 3(2):13-28, 1979.

[4] Moore, F. R., M. V. Mathews, and J. R. Pierce. "Spatialization of Sounds over Loudspeakers." In *Current Directions in Computer Music Research*. Cambridge, MIT Press, 1989.

[5] Amatriain, X., T Höllerer, J. Kuchera-Morin, and S. T. Pope. "Immersive Audio and Music in the AlloSphere". *Proc ICMC. 2007.*

[6] M. Geier, et al. "The SoundScape Renderer: A versatile software framework for spatial audio reproduction." In *Proc WFS Workshop,* Ilmenau, Germany, Sept. 2007.

[7] Pope, S. T. and C. Ramakrishnan. "The CREATE Signal Library ('Sizzle'): Design, Issues, and Applications". *Proc ICMC 2003.* See also http://FASTLabInc.com/CSL