Approaches to Immersive Audio Content Creation

Richard Foss[1], Antoine Rouget[2]

[1] *Computer Science Department, Rhodes University, Grahamstown 6140, South Africa: r.foss@ru.ac.za*

[2] *DSP4YOU Ltd., Kowloon, Hong Kong: arouget@dsp4you.com*

Abstract

This paper addresses the problem of generating an immersive sound content creation system, using as a context experience with developing a particular system, the 'Immergo' client server based system. The core components of such a system are described, with reference to standards-based and proprietary documentation. These core components are: an object model for the object audio subcomponents, the audio sample data, metadata such as position and gain to guide localization, loudspeaker configuration, renderers to render the audio content using appropriate algorithms, the final bitstream formatting, the user interface for the user to guide localization, and the system configuration. The paper contrasts current approaches and highlights the importance of standards to direct the format of the content, thereby making it universally accessible.

1. Introduction

The term 'Immersive Audio' refers to sound that appears to emanate from one or more locations around a listener. A number of computer based strategies have developed to enable the realization of this phenomenon. A consistent concept that underpins these strategies is that of 'object audio'. Object audio refers to an enhancement of the multiple audio sample files that typically make up a multichannel production. The enhancement takes the form of metadata. This is data that provides information about other data. In the case of immersive audio, the 'other data' is audio waveform data, and the metadata provides information primarily about the 3D localization of the waveform, and its loudness.

This approach is to be distinguished from a channel based approach where audio channels are pre-mixed to speakers at particular locations, and which is out of the scope of this paper.

Past years have seen the development of a number of Digital Audio Workstations (DAWs) that allow a user to create multitrack recordings, where each track comprises audio waveform data. In order to create object audio content, the Digital Audio Workstation-based content creation system must be enhanced to allow a user to direct the localization of the various track-based sound sources. An immersive audio content creation system will incorporate multiple loudspeakers, and localization is achieved by controlling parameters (such as gain) of the audio channels sent to these loudspeakers.

This paper describes the process of creating one such computer based system, known as 'Immergo'. Immergo is a client/server based system, where the server is either an Apple or Windows workstation connected to an Ethernet AVB network of audio devices, and the clients are mobile devices. The system allows a user with a mobile device to select and audio track, move this track in three dimensions, and after doing this for successive tracks, to save the movements. Figure 1 below shows the configuration of a typical Immergo system:

Fig. 1: A typical Immergo immersive sound system configuration

In describing the process of creating Immergo, the core components of an immersive content creation system will be highlighted. Furthermore, there will be a description of documentation that can guide the system creation process, some of this documentation being standards-based and other, proprietary but nevertheless in the public domain. Where appropriate, features of other content creation systems will be described, to display the range of possibilities within the necessary framework of components.

Once content has been created, it will need to be packaged and distributed for use in cinemas and on home entertainment systems on a range of distribution media. The media encoders will generate bit-streams that mirror the data within the content files, and have a particular serial format.

2. The Components

Given below is a bullet list of the components, each with a brief description of its role. This will be followed by sub-sections that describe more fully each component.

- The object model - describes the various entities within an immersive audio content creation system.
- Audio samples – the digital audio waveforms.
- Metadata - time dependent data that provides positioning, loudness and other parameters to guide the playback of the various audio samples on the loudspeakers.
- Loudspeaker configuration – the positions of the speakers.
- Renderer – utilizes metadata and loudspeaker configuration to guide the playback of audio samples through speakers, in particular appropriately modifying the audio sample playback gains of the speakers.
- User Interface – how the user interacts with the system to generate parameters such as three dimensional positioning of virtual sound sources and their relative loudness.
- System configuration – the hardware components of the system that enable the realization of user intention on multiple loudspeakers.

3. The Object Model

The object model is fundamental to an object audio-based immersive sound system in the same way that an object model is fundamental to an object oriented computer system. Indeed, the object model is typically represented using the Universal Modelling Language (UML) [1]. The metadata required for the operation of the system is derived from the object model.

Figure 2 below is a generic object model that contains core entities and the relationships between them. Thus a particular audio program (or programme), whether it be a soundtrack or a distinct music piece, will comprise a number of tracks, each track having a unique ID. The track is associated with a sequence of audio samples and also associated with one or more parameters. There are many types of parameters, but typically the parameters will include the 3D position and loudness. The 3D position indicates the location in 3D space where the audio from the track should appear to play from, from the listener's perspective. The 3D position can be in the form of Cartesian coordinates (x,y,z), or polar coordinates (azimuth, elevation, radius).

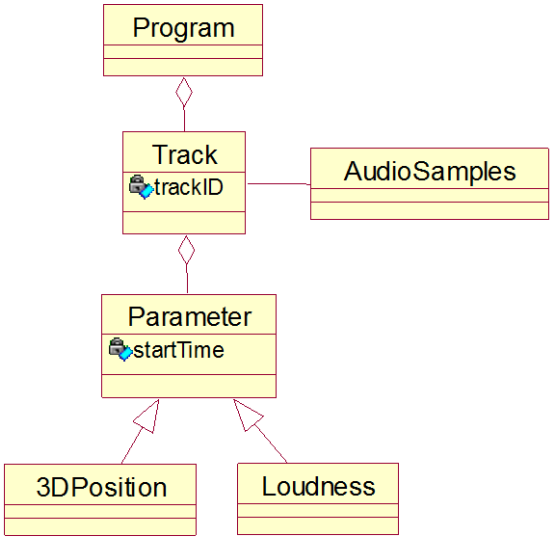

Fig. 2: Generic object model

ETSI, the European Telecommunications Standards Institute, has published a technical specification titled "MDA; Object-Based Audio Immersive Sound Metadata and Bitstream" [2]. The specification document presents an object model that incorporates the generic model, and extends it. The Program root of the model contains a number of entity objects. One of the types of Entity object is a Fragment object that incorporates the following attributes:

- Duration
- Offset
- audioEssence (audio samples)
- gain
- position

Further attributes allow for control over the 'spread' of the sound source.

EBU, the European Broadcasting Union, has published their 'TECH 3364" specification document, titled "Audio Definition Model - Metadata Specification". [3] This is currently the only specification document that comprises both an object model and accompanying metadata. The object model incorporates the generic model entities and relationships, but is more complex, because it takes account of channel and object based programs, as well as accommodating a variety of track formats.

A Broadcast Wave File (BWF) will contain audio samples for a number of tracks. The header of the BWF contains a list of track ID's, where each track ID has a reference to a track format ID and a track package ID. Theses ID's reference ID's in a metadata file whose contents are based on the object model.

The object audio model incorporates an "audioChannelFormat" object that serves to model an audio channel (sequence of audio samples) with varying formats. One of the formats is "objects", which models the dynamic

positioning of object-based audio. Each channel contains a number of "Blocks", where a Block represents a sequence of audio samples with fixed parameters in a fixed time interval. A Block incorporates the following attributes:

- start time
- duration
- position (polar coordinates)
- gain

Other attributes include the "spread" of the virtual sound source.

A number of companies have created their own immersive sound systems, with associated content creation capabilities. Dolby has the Dolby Atmos system, which provides a mix of traditional channels and audio objects. Auro Technologies created the channel-based Auro system, and DTS provides the DTS:X purely object-audio system.

The Society of Motion Picture and Television Engineers (SMPTE) has created the TC-25CSS work group to formulate standards for the files and bitstreams and hence enable interoperability amongst the files and bitstreams generated by the various companies. Work has been initiated on immersive audio metadata, but a standards document has not been released. Work in progress documents can be accessed by joining SMPTE and in particular the TC-25CSS work group

4. Audio samples and metadata

As indicated in the generic object model, there has to be a relationship between the metadata that guides the localization of the audio samples in 3D space and the audio samples themselves. The metadata will reflect the structure of the object model and is often recorded using XML. Given in listing 1 below is a simple XML layout for a few timed positions, based on the generic model:

```
<Program>
 <track TrackNumber="0"/>
 <track TrackNumber="1">
  <3DPosition startTime="0" x="0" y="0" z="0"/>
  <3DPosition startTime="260" x="-5.13" y="160"
                              z="165"/>
  <3DPosition startTime="262" x="-10.27"
                           y="165.13" z="165"/>
.
.
 <track TrackNumber="2"/>
.
.
</Program>
```

Listing 1: XML layout of timed 3D positions

Refer to [3] for a more complex object model and correspondingly complex XML code.

As indicated in the introduction, the audio samples are typically generated via DAWs and are embedded in the tracks of DAWs. In some immersive content creation systems, the metadata is generated via plugins to the DAWs. Localization control data would then be stored as DAW automation data. A difficulty here is using a single plugin to control the localization of multiple tracks.

In the case of the Dolby content creation system, the source of audio tracks and metadata is a Pro Tools system, where the metadata for each object track is provided via a plugin on the source Pro Tools track [4].

Dolby describe their complete package of audio and metadata as a 'mix' [5]. A Dolby mix comprises:

- pre-mixed channels (bed tracks) – panning is incorporated in these channels,
- audio 'objects' (object tracks), which are mono or stereo sound tracks whose panning will be controlled by metadata,
- metadata, which includes time synchronized panning data.

The Pro Tools-based Dolby Atmos content creation system will create a 'print master'. The print master incorporates:

- Ten mono .wav files, which are bed tracks,
- One .prm file and one .wav file per object track. The .prm file contains the panner data – positions and times that are created in the plugin for the track.
- One dub_out.rpl file. This is an xml file that contains information about all the .wav files.

The bed tracks and the object tracks are created in Protools. This can be contrasted with:

1. the EBU TECH 3364 approach, where a BWF file contains all audio samples, and the header of the BWF file references tracks in a separate XML metadata file, and
2. the Immergo approach, where there are multiple.wav files (for each track in the DAW), and a single XML metadata file with implicit relationships.

In all cases, there is a need for:

1. Audio samples
2. Metadata
3. Relationships between 1 and 2

Most important is to create a standard that lays out file formats for the above three components, and that is widely adhered to.

5. Loudspeaker configuration

For the renderer to perform the task of sound localization, it has to know the location of the loudspeakers. Distinguishing

features of immersive sound systems are their enhanced number of loudspeakers, the range of loudspeaker configurations, and particularly the fact that the loudspeakers are at varying height levels around the listener.

Three-dimensional coordinate values need to be provided for each speaker, in relation to a pre-determined origin. This origin could, for example be the bottom left front corner of the room, or the center of the room.

In the case of the Dolby Atmos content creation system, a room configuration in the form of a ".dac" file is loaded into the Rendering and Mastering Unit (RMU).

The Immergo system requires an XML file with the speaker locations. An excerpt from the speaker configuration file is given Listing 2 below:

```
<config>
   <roomheight> 320 </roomheight>
      <speaker number = "1" xpos = "0" ypos = "0" zpos = "100"
      </speaker>
      <speaker number = "2" xpos = "-104" ypos = "0" zpos = "100"
      </speaker>
      <speaker number = "3" xpos = "104" ypos = "0" zpos = "100"
      </speaker>
      <speaker number = "4" xpos = "-104" ypos = "0" zpos = "230"
      </speaker>
.
.
.
</config>
```

Listing 2: XML for speaker configuration

Manual measurement of speaker positions is tedious and error prone, and work has commenced on the automation of this process as part of the Immergo system.

6. Renderer

Since the user of a content creation system needs to hear the result of localization requests, it is important that these requests are 'rendered' to loudspeakers in the system configuration. The task of a renderer is to utilize time dependent localization requests for sound sources, and in the context of the loudspeaker configuration, ensure that these sound sources are correctly positioned in three dimensional space. Figure 3 below gives a diagrammatic representation of the role of the renderer in the context of the Immergo immersive sound system.

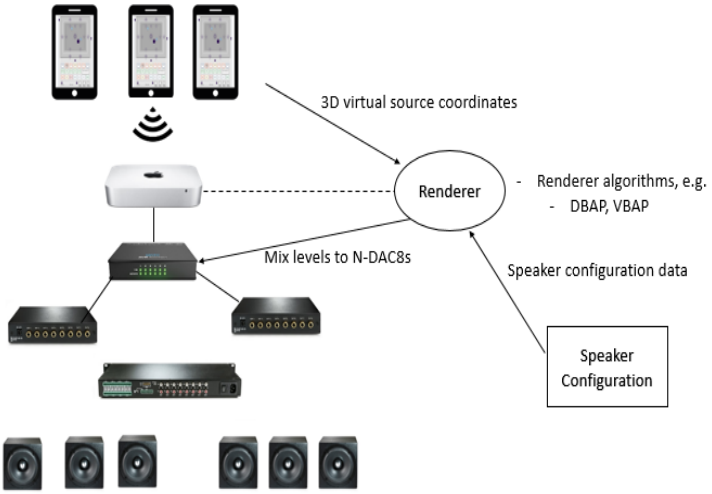

Fig. 3: The role of the Renderer in the Immergo immersive sound system

In order to perform its task, the renderer has to use some mechanism to create the appropriate parameters, in particular gains, for the control of output audio channels. Historically, Vector Based Amplitude Panning (VBAP) [11] and Ambisonics [14] have been used to localize sound sources. However, recently Distance Based Amplitude Panning (DBAP) has been proposed as a mechanism that avoids some of the limitations of VBAP and Ambisonics, in particular the restrictions on listener and loudspeaker positioning [13].

Both DBAP and VBAP, as their names imply, are amplitude panning techniques that apply amplitude panning to the various tracks in a multi-track recording, and thereby achieve localization. They extend stereophonic panning, the popular method for two dimensional amplitude panning.

Ambisonics does not apply amplitude panning to multiple audio channels, but rather decodes a speaker-independent representation of the soundfield, called B-format, to a particular speaker setup. The EBU 3364 specification does allow for the inclusion of such scene based channels [3].

VBAP requires a triangulation process to determine triplets of loudspeakers. A virtual sound source position is used to determine its associated triplet. These loudspeakers and their direction/distance from the listener position are then used as a vector base for a virtual source's direction vector. Appropriate gains for the real source are applied to each loudspeaker of the triplet. The reference renderer in the MDA specification uses VBAP to map object waveforms to loudspeaker outputs [2].

DBAP in contrast simply determines the gain to be applied for each channel, for each speaker, by assuming that it is inversely proportional to the distance between the speaker and the desired virtual source position for the channel. The Immergo system allows a choice of rendering techniques, either DBAP or VBAP.

A specification that determines the files generated in the process of content creation should not necessarily prescribe a rendering mechanism to be used. However, it should have the capability to indicate what mechanism that was used in the

content creation process. This will enable playback localization to match the localization at content creation time.

The ISO/IEC 23008-3 specification, which specifies how three dimensional audio signals are transmitted and rendered, is specific about the rendering mechanism to be used, in this case VBAP [15]. It would then be appropriate that the same rendering mechanism is used at content creation time.

7. User Interface

A user interface that is both intuitive and powerful is an essential component of an immersive content creation system. The user interface should:

1. Enable the selection of any localizable track within a multi-track arrangement.
2. Enable the placement of a selected track at any position within the 3D space of a room.
3. Allow for transport control of the DAW multitrack source.
4. Allow for the storage and retrieval of time-based localization sequences.
5. During playback allow for the visual display of the 3D movements of tracks within the room.

As indicated previously, the Dolby content creation system takes the form of a plugin associated with each Pro Tools track that is to be localized [4]. Thus track selection is implicit and the transport control within Pro Tools is utilized. There is a two dimensional display of the room and speakers, and elevation can be enabled by selecting 'elevation mode', thereby enabling movement along the z axis. Localization control is via a mouse or JL Cooper joystick.

DTS also have a content creation tool that creates files in their open MDA format. It is termed a DAW tool, so presumably is a plugin, allowing for localization of a particular track. The interface is interesting - it comprises a number of concentric circles with speaker representations overlaying the circles. Mouse movement along the outermost circle will cause the sound source to be localized at corresponding positions at the level of the lowest height speakers. Movement towards the inner circles will raise the elevation of the sound source. This is a feasible approach when using a rendering mechanism such as VBAP where the virtual sound sources are confined to a sphere around the listener. The Immergo content creation user interface shown in Figure 4 below has track selection buttons to select a localizable track in a DAW. Since the interface is not a DAW plugin, there is no restriction on the type of DAW. The system runs within a browser on a mobile device. Placement of a selected track at a 3D position can be done via a mix of touch screen control and mobile device orientation. By selecting appropriate checkboxes, sound source positioning in all three dimensions can be performed via appropriate mobile device orientation. Typically two dimensional control is via the touch screen with the third dimension acquired by tilting the mobile device. A mobile rectangle gives visual feedback regarding elevation. Currently the view is from above the room, but alternative views will be provided.

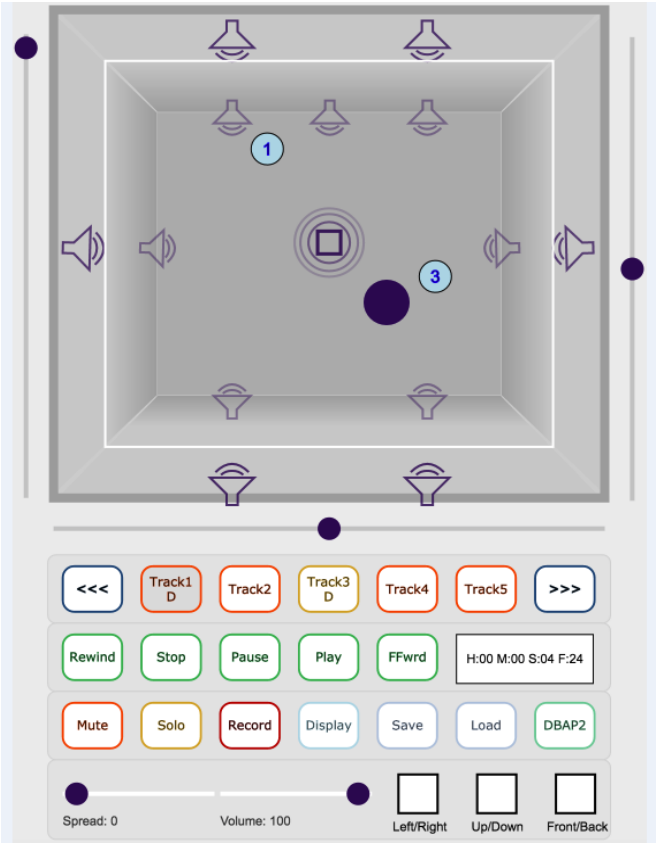

Fig. 4 – The Immergo user interface

DAW Transport control is provided on the second row of buttons. Time code is continuously displayed while the DAW is in playback or recording mode. Time-based localization sequences can be stored and retrieved. During playback/recording previously recorded tracks will be displayed as moving circles with their track numbers.

In the past few years there have been a number of documented approaches to the control of three dimensional localization. At Rhodes University gesture recognition techniques were successfully tested [8], and at the BBC a haptic feedback device was used for sound source control [7]. There is no call for standardization of user interfaces, and indeed the range of emerging techniques serves to enhance the field of spatial audio.

8. System Configuration

There are currently two distinct system configurations for content creation:

1. A plugin approach, where the content creation application is tightly bound to the underlying DAW.
2. A client/server approach, where a server application runs alongside a DAW, and serves a web browser with client code.

Examples of the first configuration have already been described. A further example of this approach is the "Spatial Audio Designer" from New Audio Technology [16]. Fraunhofer Institute have created and documented an example of the second approach [6]. The Immergo system also uses the second configuration [9].

Both the Fraunhofer and Immergo systems allow multiple users to control aspects of multitrack localization content creation at the same time. Each user can perform this control remotely on a mobile device of their choice using a browser of their choice. Unlike the first system configuration, these two systems have to ensure that there is a common clock that can time localization events in the server application. Both systems use time code for this purpose.

Regardless of the system configuration, a content creation system will need some form of processing to perform appropriate mixing of multiple channels before they are output to loudspeakers. Delay processing as well as other signal processing such as filtering will often be required. There are two possible approaches to this processing requirement:

1. A single powerful processor box that incorporates a large (n x m) matrix mixer and other signal processing functions.

2. A distributed approach, whereby multiple channels of audio are distributed to loudspeaker units that each incorporate processors. In this case the mixer matrix in each loudspeaker will be (n x 1)

The Immergo system uses the second approach since it enables progressive enhancement of a system, and does away with the need to invest immediately in a large and expensive processor box. Figure 4 below is an outline diagram of the Immergo system configuration.

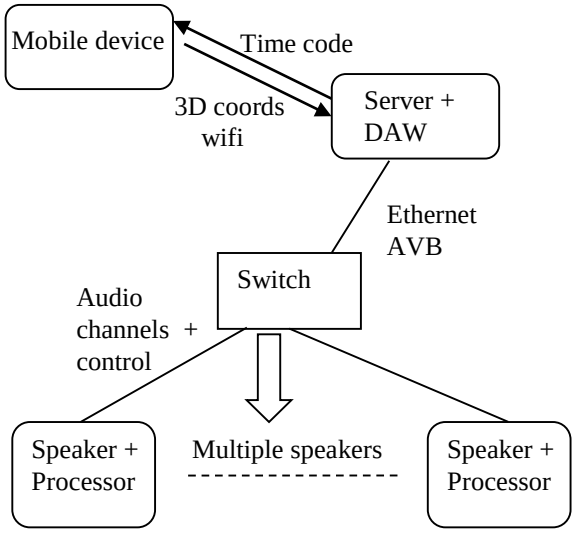

Fig. 4 - Immergo system configuration

In this case the network used is Ethernet AVB, and the speaker processors are Ethernet AVB endpoints. This is in contrast to Figure 1, where two Ethernet AVB multi-output interface boxes performed the required audio processing. Both configurations are possible, each having their advantages for particular applications.

The mobile device transmits localization information to the server over a wireless network. The server application uses DBAP or VBAP algorithms to determine the mix gains for each channel on each loudspeaker. These parameters are transmitted to the speakers, where the mixing takes place alongside delay and filter processing.

As with user interface approaches, there is no need to standardize on system configurations, and the diversity enhances the field of spatial audio.

9. Conclusion

The application of the object audio concept to immersive sound promises to enhance multi-channel localization control, and hence the listener experience in many contexts. Whatever the context, there is a need to create content that incorporates localization information. If this content is to be used widely and in many contexts, then the files produced have to follow a standard format. This paper has highlighted the need for such standardization, and indicated the various bodies that are involved in the standards process. The paper has also described the many components that constitute a content creation system. While some components, such as the audio sample and metadata files require standardization, other components such as the user interfaces enhance the field with their diversity.

10. References

[1] Object Management Group: OMG Unified Modeling Language version 2.5. March 2015.

[2] ETSI: ETSI TS 103 223 v1.1.1 – MDA; Object-Based Audio Immersive Sound Metadata and Bitstream. April 2015.

[3] European Broadcasting Union: TECH 3364 Audio Definition Model – Metadata Specification version 1.0. January 2014.

[4] Dolby Laboratories, Inc.: Authoring for Dolby Atmos Cinema Sound Manual. issue 1, 2013.

[5] Dolby Laboratories, Inc.: Dolby® Atmos® Next-Generation Audio for Cinema – white paper. issue 3, 2014.

[6] G. Gatzsche and C. Sladeczek,: A Flexible system Architecture for Collaborative Sound Engineering in Object-Based Audio Environments. Audio Engineering Society 136th convention, April 2014.

[7] F. Melchior, C. Pike, M. Brooks, and S. Grace: On the use of a haptic feedback device for sound source control in spatial audio systems. Audio Engineering Society 136th convention, May 2013.

[8] M. Hedges, R. Foss: Utilizing gesture recognition and Ethernet AVB for distributed surround sound control. Audio Engineering Society 136th convention, October 2014.

[9] R. Foss, A. Rouget: A Method of Positioning an Output Element within a Three Dimensional Environment. PCT International patent application No. PCT/IB2016/052117.

[10] IEEE: IEEE Standard for Device Discovery, Connection Management, and Control Protocol for IEEE 1722 Based Devices. Document Standard, IEEE Std. 1722.1, 2013.

[11] V. Pulkki: Virtual Sound Source Positioning using Vector Based Amplitude Panning. Journal of the Audio Engineering Society, vol. 45, no. 6, pp. 456-466,1997.

[12] T. Lossius, P. Baltazar, and T. de la Hogue: DBAP - Distance-Based Amplitude Panning. International Computer Music Conference (ICMC). Montreal, 2009.

[13] D. Kostadinov, J. Reiss, and V. Mladenov: Evaluation of Distance Based Amplitude Panning for Spatial Audio. ICASSP, 2010

[14] R. K. Furness: Ambisonics - An Overview. Audio Engineering Society 8th International Conference, Washington, D.C., 1990

[15] ISO/IEC 23008-3: Information technology - High efficiency coding and media delivery in heterogeneous environments - Part 3: 3D audio. October 2015.

[16] This is the reference to the New Audio Technology Spatial Audio Designer information page,
URL:
http://www.newaudiotechnology.com/en/products/spatial-audio-designer/