Notes for lecture/ Zack Settel, McGill University

Sound source localization and its use in multimedia applications

Introduction

With the arrival of real-time binaural or "3D" digital audio processing, techniques for acoustic space modeling and sound source spatialization have become of great interest for implementation in various multimedia applications, ranging from CD-ROM authoring to video conferencing. In surround sound applications, where sound source localization is a primary objective, various digital signal processing (DSP) techniques can be employed to heighten the sense of source localization, often creating a stronger link between sound and image. In several multimedia applications, the degree of live interactivity can play a crucial role in the success of the application. For example, the underlying software in video games, acoustic space simulators and air traffic monitoring systems (using auditory display) must render audio objects in a real-time acoustically modeled space, based on varying position data supplied to these programs by the user(s). This paper will focus on sound source localization techniques and their use in various multimedia applications, and will examine key issues in implementation.

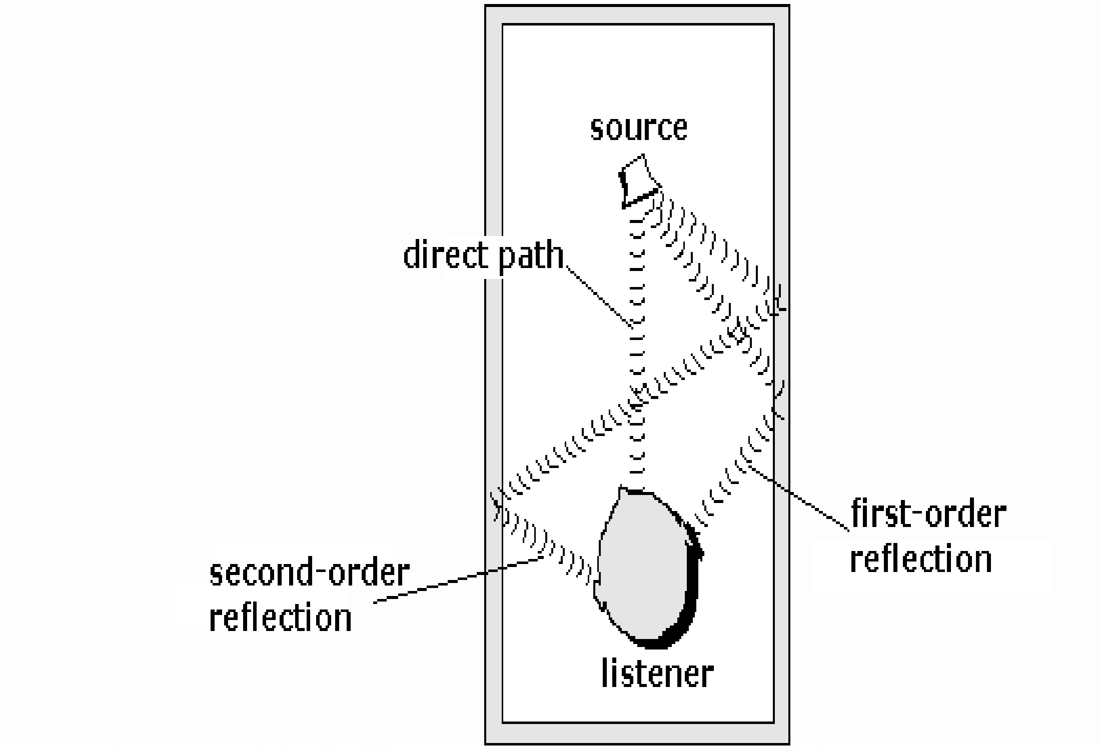
What is sound source localization?

Today I will present two approaches to sound source localization, which make use of virtual sound source location using binaural source imaging techniques (3-D audio), and/or physical sound source location using multiple speaker fields (surround sound). I will discuss the techniques involved in the design, implementation and use of these systems, and then go on to discuss three systems, which make varied use of these techniques.

Technique 1: Localization using Virtual: Sound source Imaging

"3D audio", a recently coined term, refers to technologies that are based on the principles of binaural human hearing. Interactive 3D audio allows the user to dynamically specify the position of sounds sources in the three-dimensional space surrounding the listener. Recent developments in 3D audio among computer game and multimedia companies have given rise to popular interfaces, such as MicroSoft's DirectSound 3D, or Aureal.

Technique 1: Localization using Virtual: Sound source Imaging



Typical sound field with a source, environment, and listener (*image courtesy Aureal corp.*)

3D audio systems typically divide into three parts:

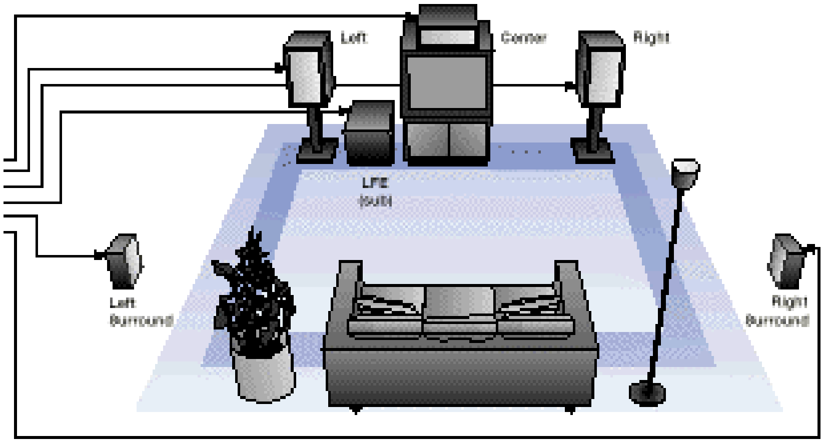
•**the sound source**: any signal

•**the acoustic environment**: A system modeling the source signal's propagation to the listerers' ears.

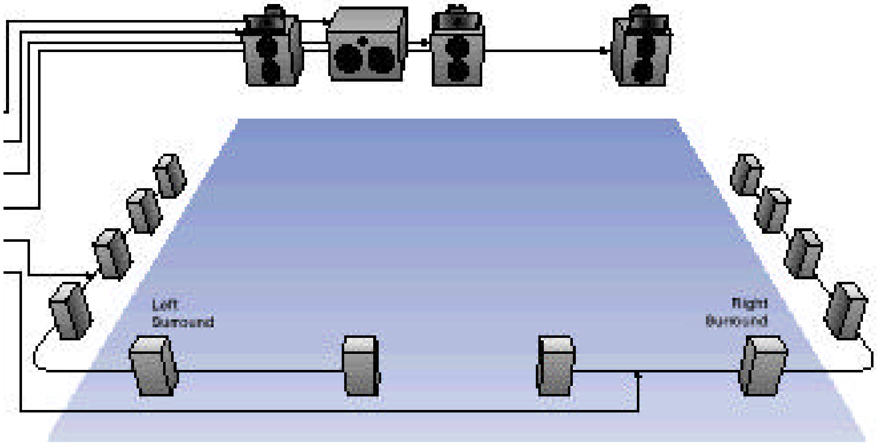
•**the listener** (a pair of ears): A system modeling the way signals behave at the ear, to provide the listener with acoustic cues to interpret information about the location of the sound sources and the environment.

(*images courtesy Aureal corp.*)

Technique 2: Localization using placement of sound in a multiple speaker field



typical 5.1 "Surround Sound" configuration



example of a Cinema "Surround Sound" configuration

(images courtesy Dolby Laboratories Inc.)

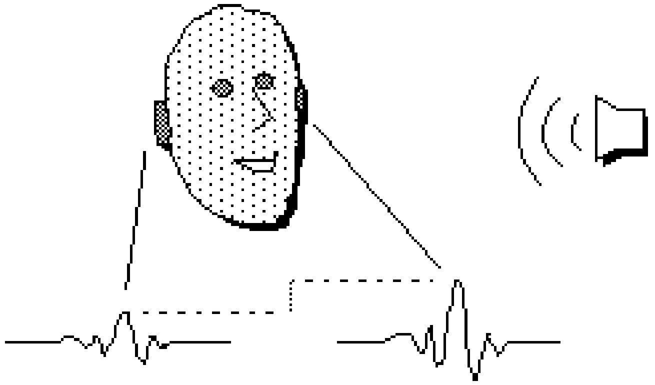
Human Hearing

The auditory system consists of two ears and a brain, which uses cues embedded in the signals it receives from the ears to locate sound sources and determine environment quality.

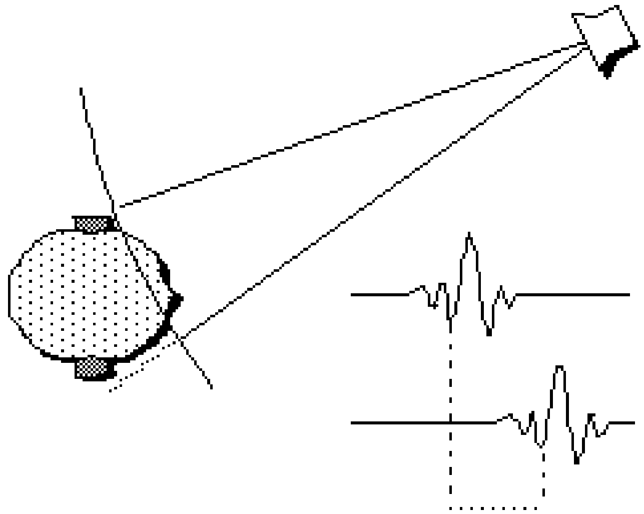
The following cues are widely recognized as being important in this determination:

1. Primary Localization Cues:

- Interaural intensity difference (**IID**)
- Interaural time difference (**ITD**)



interaural intensity difference (IID)

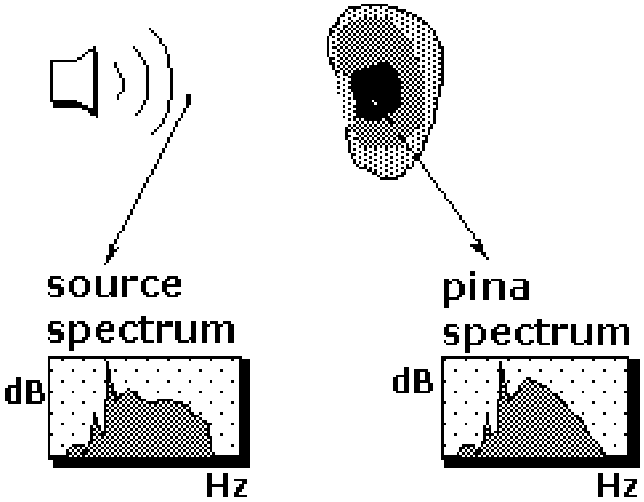


interaural time difference (ITD)

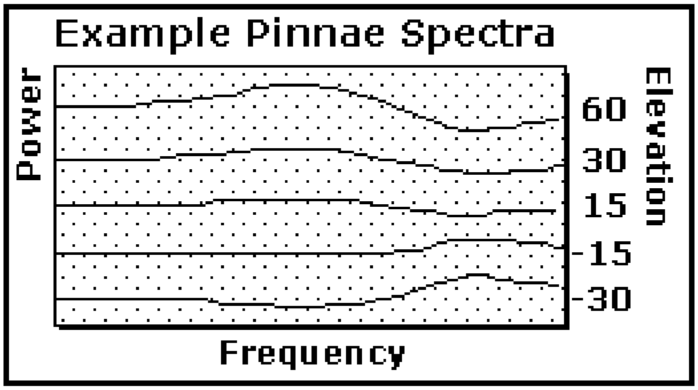
(images courtesy Aureal corp.)

2. Physical structure of the outer ear:

The pinnae are the key to accurately localizing sounds in space. A pinna, acting as a filter, boosts or cuts mid- and high-frequency energy of a sound, depending on the location of the sound source.



spectrum differences between original and received signal

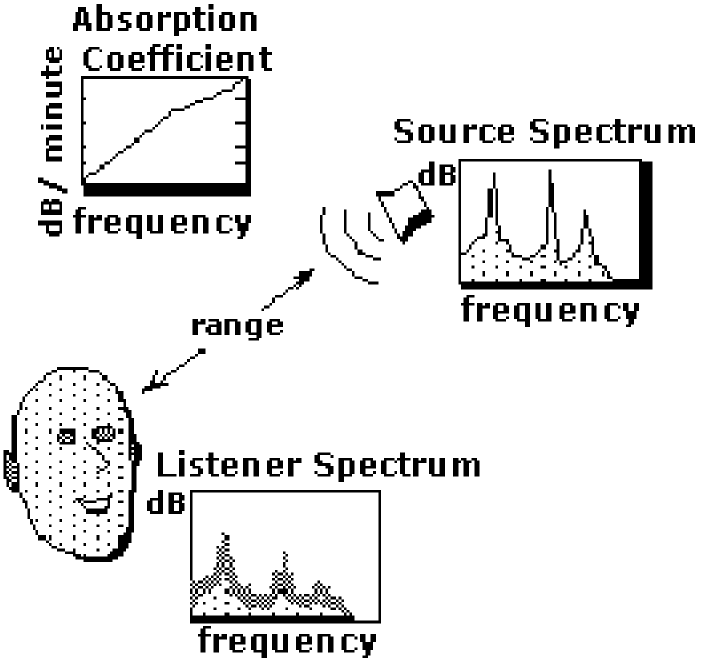


Alterations to frequency content of sounds arriving at the pinna from varying elevations

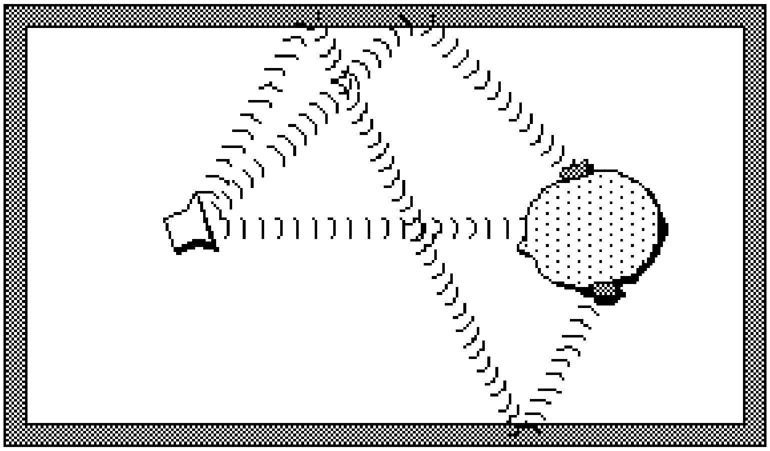
(images courtesy Aureal corp.)

3. Reflections, Propagation Effects, Proximity

Determining the quality of the environment and distance to source.



source attenuation and absorption

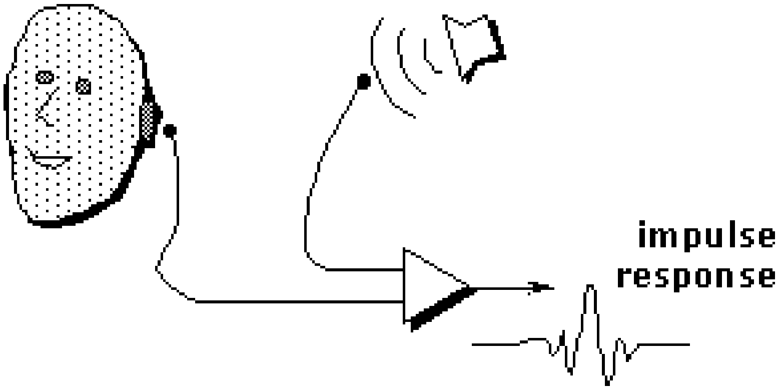


direct sound and first and second order reflections

(images courtesy Aureal corp.)

Head-Related Transfer Functions (HRTF)

A HRTF is often realized as a pair of audio filters (one for each ear) that model the frequency responses within a particular listener’s left and right ear canals, to sounds which originate from specific locations. Filter settings (known as "ear prints") are gathered and stored, one set for each specified sound source location.



acquiring HRTF data for a particular listener

(*image courtesy Aureal corp.*)

Interactivity:

The link between visual, and/or movement and sound is compelling.
Example:
Flight simulators, virtual attractions (Epcot Center etc.)

Sound de-synchronized to image, can falsely represent action times, within tolerances of hundreds of milliseconds.

Real-time responses of interactive devices ranging from high-performance automobiles to electronic musical instruments dictate the degree of transparency (immersion) of the experience.

Challenge: to find meaningful mappings of gesture to system parameters

Interactive localization:

- Two-dimensional controllers for surround sound mixing.
- Multi-dimensional gaming interface (navigation / position)
- Movement Tracking (teleconferencing, air traffic control etc.)
- Listener head position tracking

The audio cues for localization change dramatically when the listener tilts or rotates their head. Ambiguities in sound source localization are often resolved by the listener's moving their head to "have a better look". A faint, low sound could be either in front or back of us, so the listener briefly turns their head to the left; if the sound is now off to the right, it is in front, otherwise it is the back. Using sensors to track the listener's head position, the sound source position can be updated accordingly, adding a greatly heightened sense of "immersion" to the experience.

Featured systems

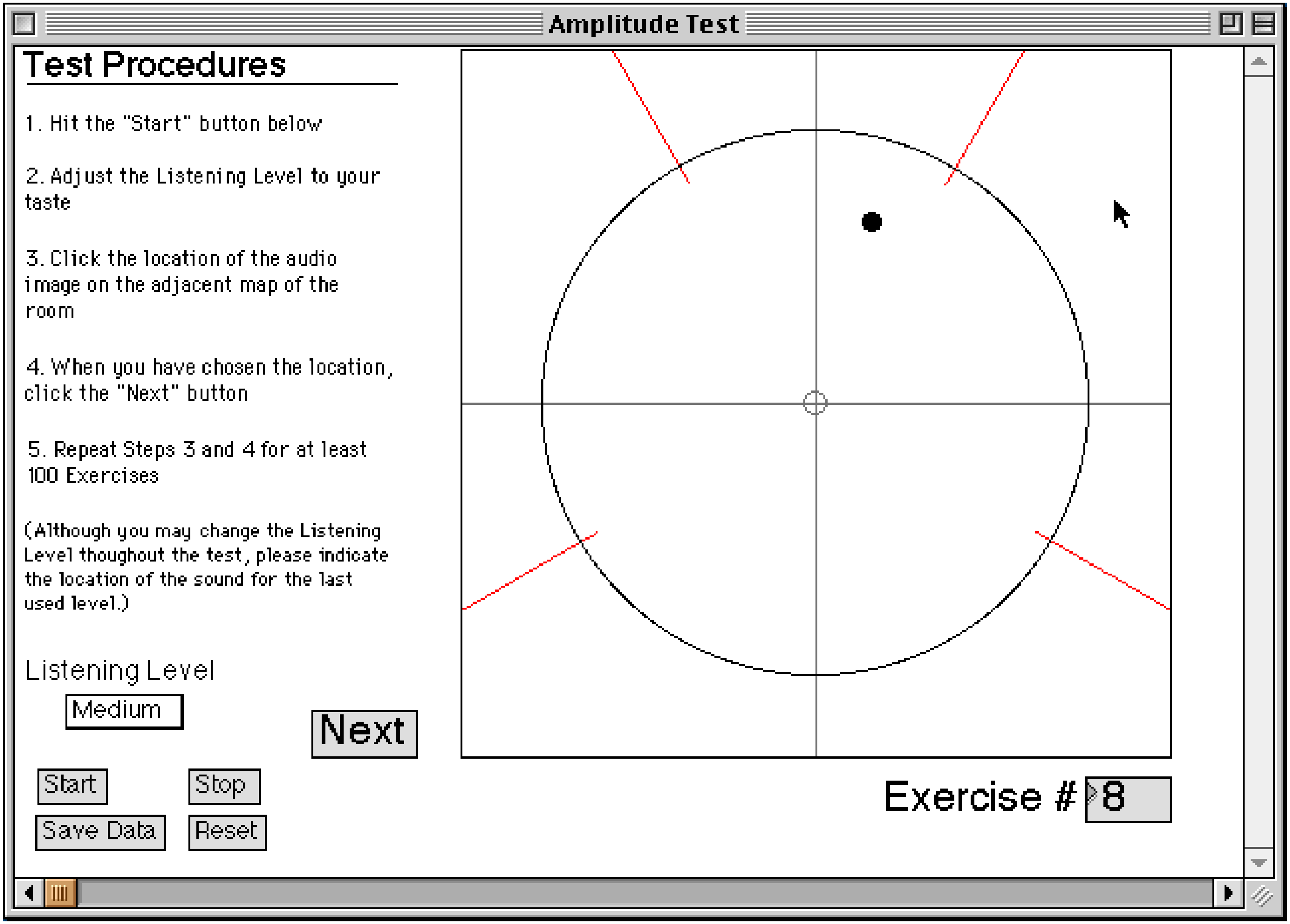
Three systems are presented, each one focusing on a different aspect in the practice of sound source localization. All three were developed in the same programming environment (Max/MSP, from ISPW/FTS). I will start by discussing a research-based system, in which test subjects assist in the empirical derivation of sound source localization parameters. Then a system with an extended interface for multi-channel surround sound localization will be presented. Finally, a system integrating 3D audio and surround will be shown.

System 1. A 5-channel Audio Image Location and Acoustic Space Simulator,

Developed by Geoff Martin
Multichannel Audio Research Lab
Music, Media and Technology Group
McGill University Faculty of Music

Important concepts:

Research System in its early stages: work is currently focused on a series of two listening tests, the intention of which is to find derive localization data for a “standard” 5-speaker discrete multichannel playback system with loudspeakers positioned at 0°, ±30° and ±120° and again, equidistant to the listener. Test subjects were presented with program material played through two adjacent loudspeakers chosen randomly by the testing software, which was developed by Geoff Martin of the McGill Music Media and Technology Group. Two similar tests were performed, the first to investigate the relationship between interchannel amplitude differences and the perceived image direction in the horizontal plane; the second to correlate interchannel time differences with the perceived image direction.



Test Procedures

1. Hit the "Start" button below

2. Adjust the Listening Level to your taste

3. Click the location of the audio image on the adjacent map of the room

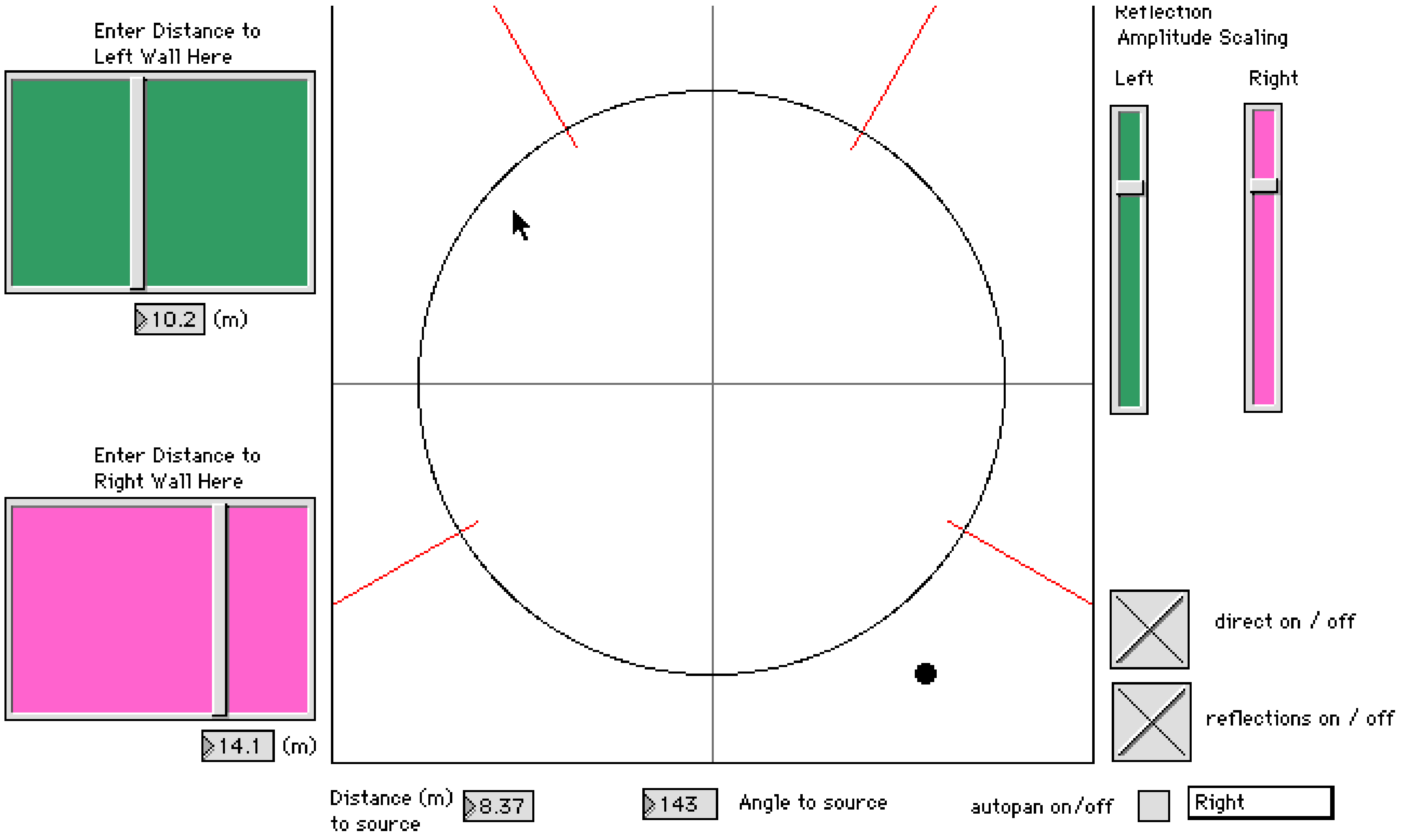
4. When you have chosen the location, click the "Next" button

5. Repeat Steps 3 and 4 for at least 100 Exercises

(Although you may change the Listening Level thoughout the test, please indicate the location of the sound for the last used level.)

Listening Level: Medium

Exercise # 8



Enter Distance to Left Wall Here

10.2 (m)

Enter Distance to Right Wall Here

14.1 (m)

Reflection Amplitude Scaling

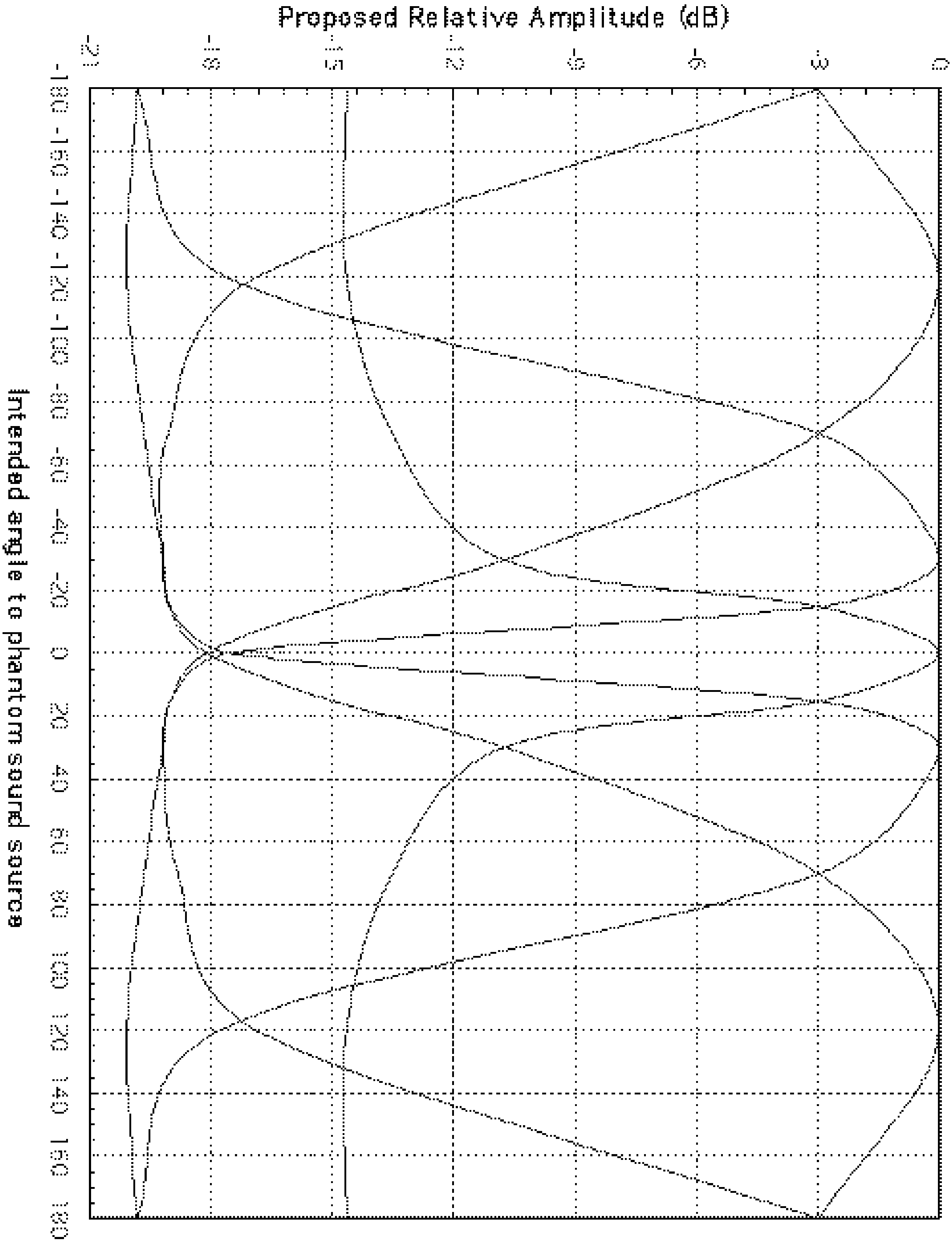
☒ direct on / off

☒ reflections on / off

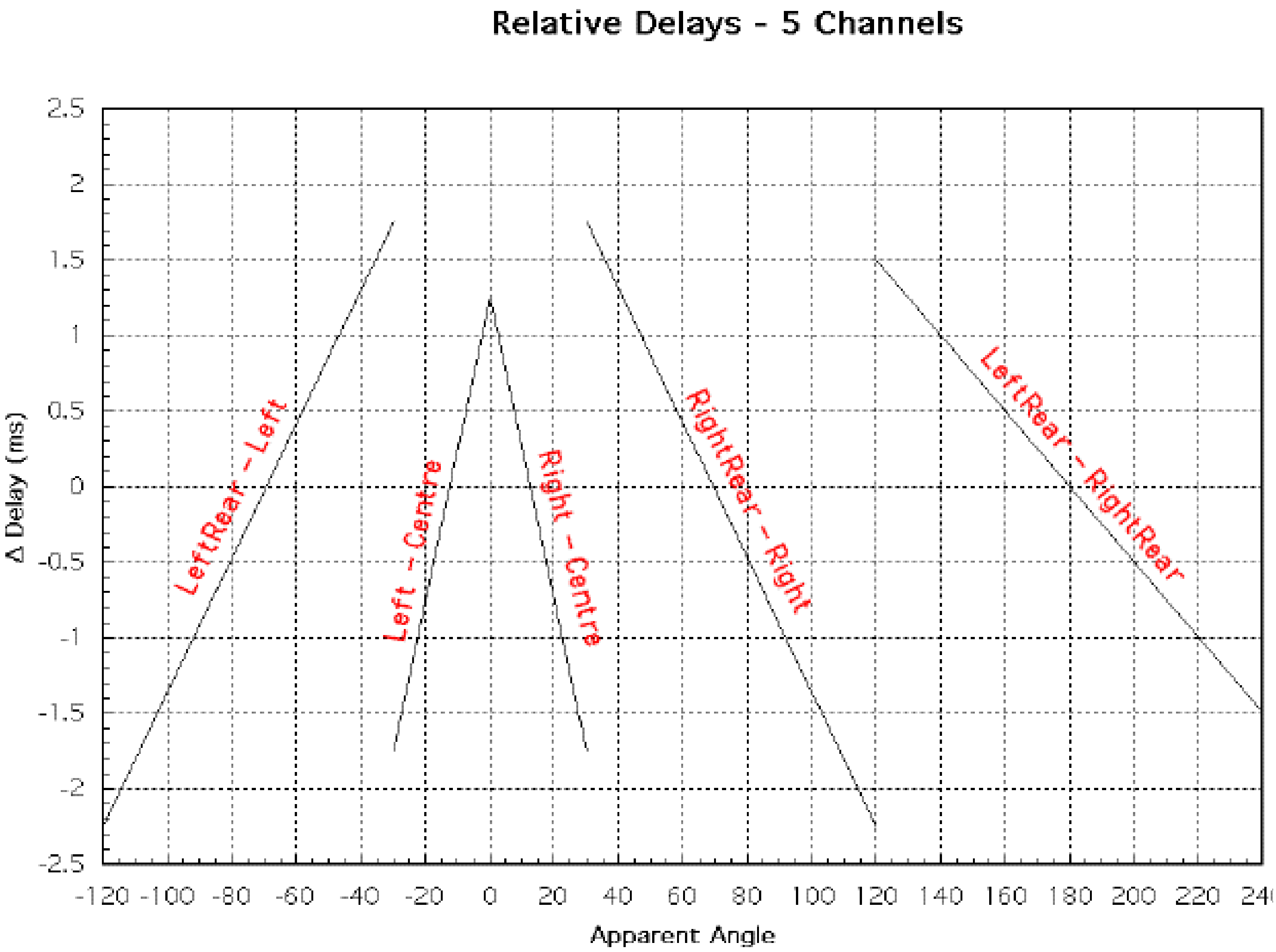
Distance (m) to source 8.37

143 Angle to source

autopan on/off □ Right


Proposed Relative Amplitude (dB)
-21
-18
-15
-12
-9
-6
-3
0
-180
-160
-140
-120
-100
-80
-60
-40
-20
0
20
40
60
80
100
120
140
160
180
Intended angle to phantom sound source

(images courtesy of McGill Music Media and Technology Group)


Relative Delays - 5 Channels
Δ Delay (ms)
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-120
-100
-80
-60
-40
-20
0
20
40
60
80
100
120
140
160
180
200
220
24(
Apparent Angle
LeftRear - Left
Left - Centre
Right - Centre
RightRear - Right
LeftRear - RightRear

(*images courtesy of McGill Music Media and Technology Group*)

System 2. The Localizer, Zeep.com

Based on Interaural intensity difference (**IID**) with the addition of reverb shadowing for the Interaural time differences (**ITD**).

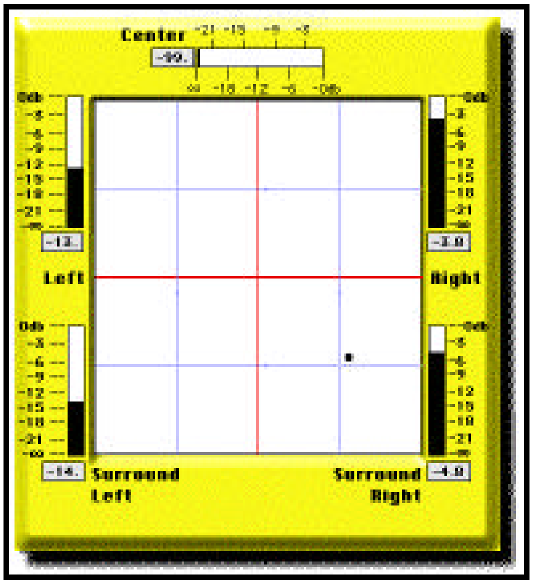
Important concepts:

1. Deals with multiple (up to 40) surround-sound inputs, using Master/Slave concept for linked movement of multiple sources. Linked movement modes include: Normal, Invert X, Invert Y, Invert XY.

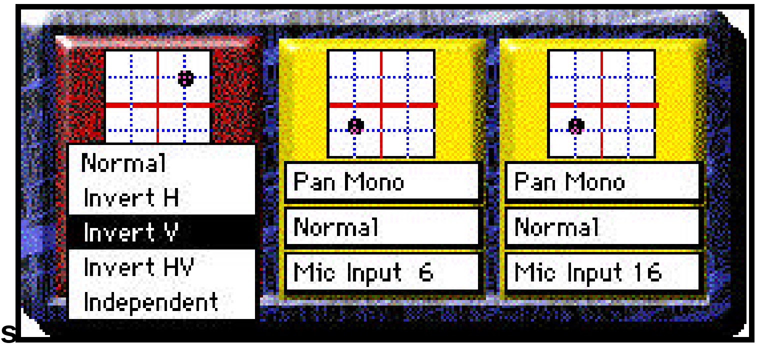
2. Provides external control of sound source location via MIDI.

3. Provides linked movement for reverberated or delayed sources, allowing for interaural time differences in the surround field.

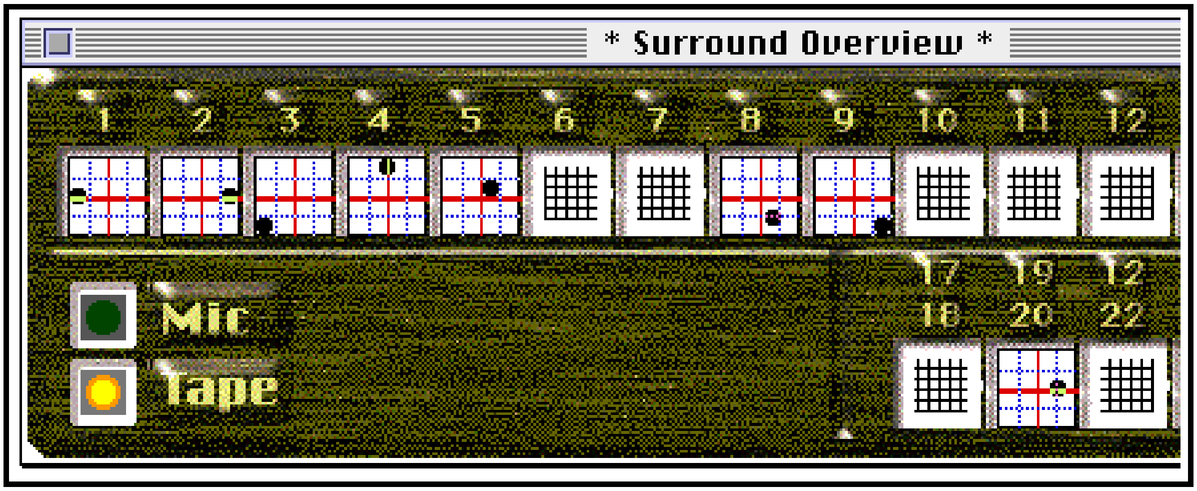
Multi-channel Localization interface for 5.1 Surround Sound



master source location at point (x,y)



slave linked sources linked to master



view of independent source locations

(images courtesy of Zeep.com)

System 3. The Spatializer, from IRCAM, France (j-m. Jot)

Spat is a 3-D audio rendering and artificial reverberation software allowing reconstruction of three-dimensional virtual sound scenes by processing and mixing mono or stereo source signals. The modularity and configurability make the software adaptable to a variety of application contexts, including musical composition in the studio, concert performance, sound reinforcement, post-production of recordings and soundtracks, multimedia or virtual reality systems. Spat runs in IRCAM's FTS environment, as well as in Cycling-74's MSP environment, from which the transparancies for today's talk were taken.

Spat receives a mono or stereo source signal and applies directional panning and reverberation processing to produce a multi-channel output to feed headphones or various loudspeaker setups in a wide variety of formats.

A typical application of Spatializer consists of associating several Spat~ objects in parallel, each receiving a different mono or stereo source signal, and mixing their outputs to render a complex 3-D sound scene.

KEY FEATURES

- A control interface is proposed which allows to specify the desired effect using perceptual terms (derived from psychacoustic research carried out at IRCAM) rather than technological terms. Since each perceptual attribute is linked to an objectively measurable criterion of the transformation of the sound, this control interface allows to imitate the acoustics of an existing room. It then allows to interpolate or extrapolate continuously towards a different acoustic quality, going through natural-sounding transformations.

- integrating directional panning and artificial reverberation in a single processing module providing a high-level control interface

- configurability according to the output format and the rendering technique (can drive headphones or 2 to 8 louspeakers, using binaural techniques, conventional stereo techniques, intensity panning, B-format and Ambisonic decoding, Dolby-compatible encoding)

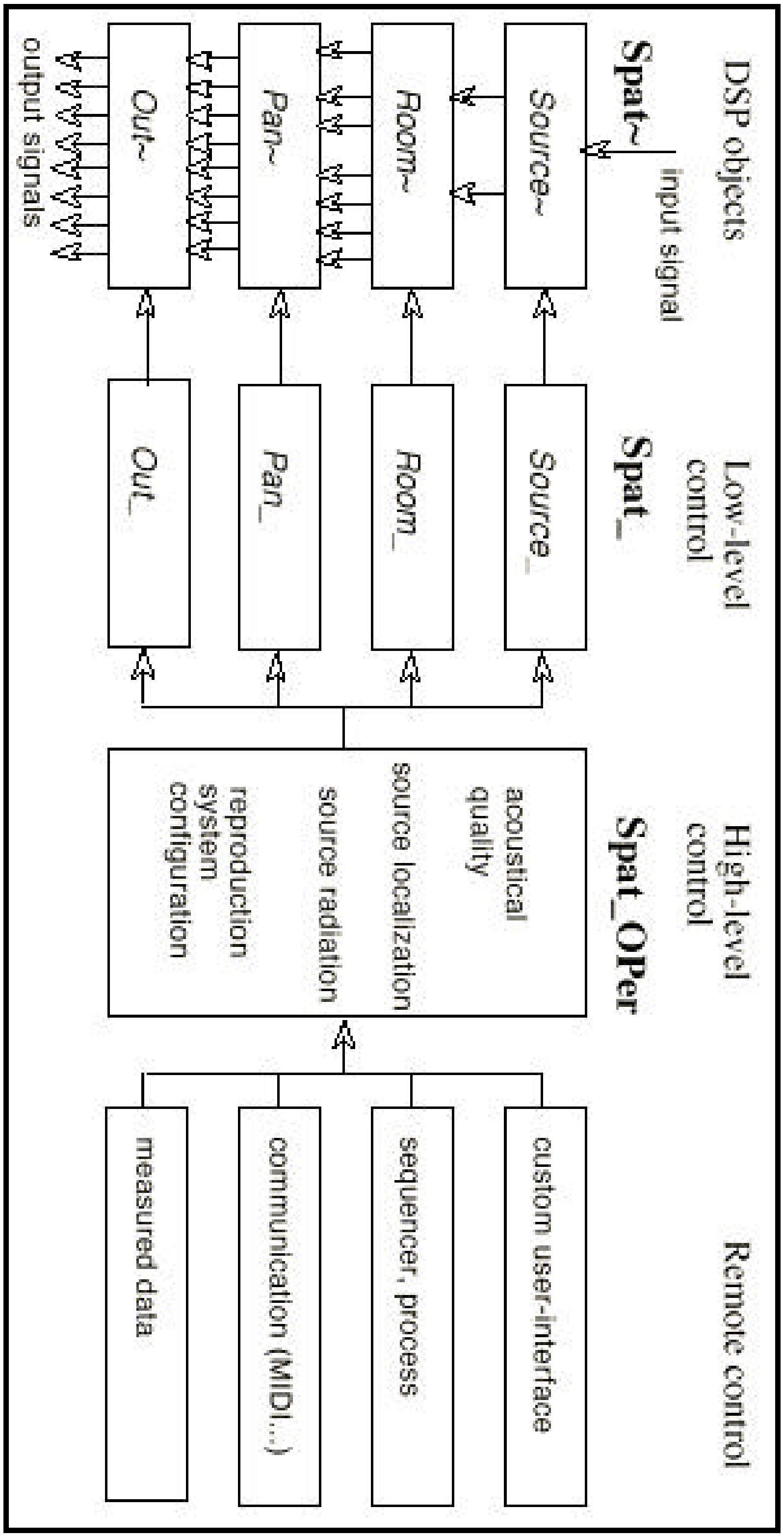
- efficient and natural-sounding artificial reverberation algorithms, and scalability of the reverberation model allowing to adapt the fidelity / comptational cost trade-off according to the audio context and the available resources

- efficient implementation of binaural synthesis (HRTF filters) for 3-D spatialization over headphones and 2 or 4 loudspeakers

- high-level control interface providing a set of control parameters which is independent of the configuration of the output format and of the reverberation model.

- Transaural reproduction: A two-channel binaural signal (obtained by binaural recording of a real sound field or synthesized electronically by binaural processing or mixing) can be converted into a transaural signal (to be reproduced over two loudspeakers). This is necessary in order to compensate for the effect of the 'cross-talk' acoustic paths (left speaker to right ear, right speaker to left ear). This post-processing of a binaural signal will allow to preserve, when the transaural output is reproduecd over two loudspeakers, the original localizations of sounds as encoded in the binaural input signal.

System 3. The Spatializer, from IRCAM, France (j-m. Jot)

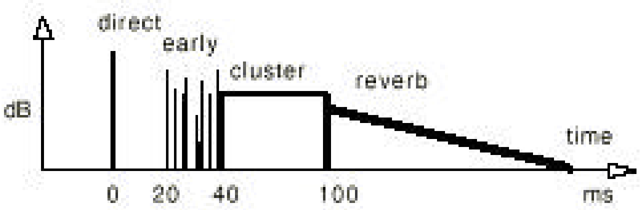


(*image courtesy of IRCAM*)

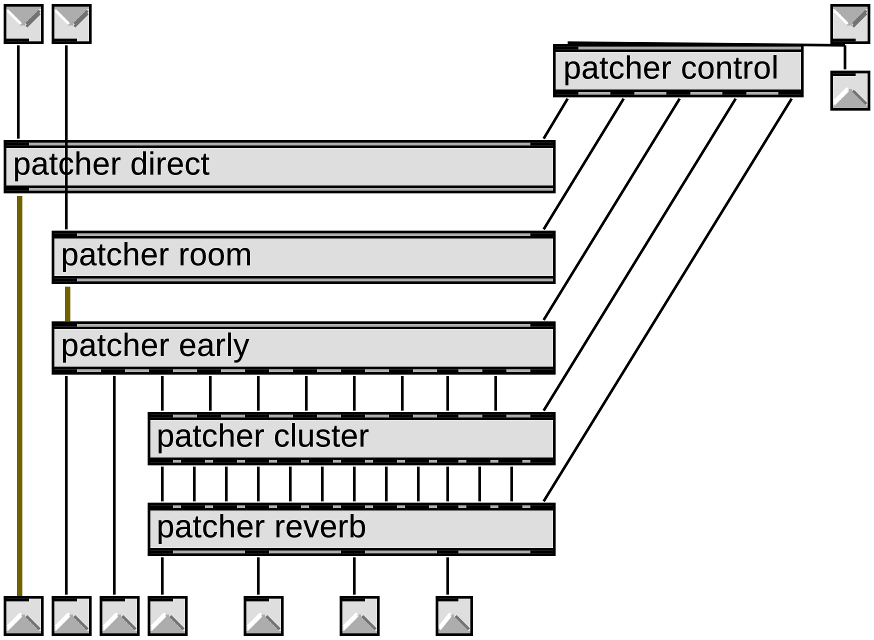
Source: realizes a “pre-delay” in order to reproduce, if necessary, the time lags existing between the signals coming from several sound sources situated at different distances from the listener. A continuous variation of this pre-delay naturally reproduces the Doppler effect (apparent pitch shift) associated to the movement of a particular source. Low-pass filtering to reproduce the effect of air absorption is also included in **Source.**

Room~: Signal Processing and low-level control The response of Room~ is divided into four time sections: ***direct*** The direct sound is taken as the time reference (0 ms) for the description of the artificial room effect that follows it. It is sent to the *center* output of Room~. ***early*** This section contains the discrete early reflections, shared between the two *sides* outputs of Room~. The date and intensity of each reflection can be controlled individually. ***cluster*** This section contains a denser pattern of diffuse later reflections which are equally shared between the four surround ouputs. ***reverb*** This section contains the late diffuse reverberation, divided into four uncorrelated signals of equal energy sent to the surround outputs. The

about Room~



room~: reflections vs time



room~: implementation structure

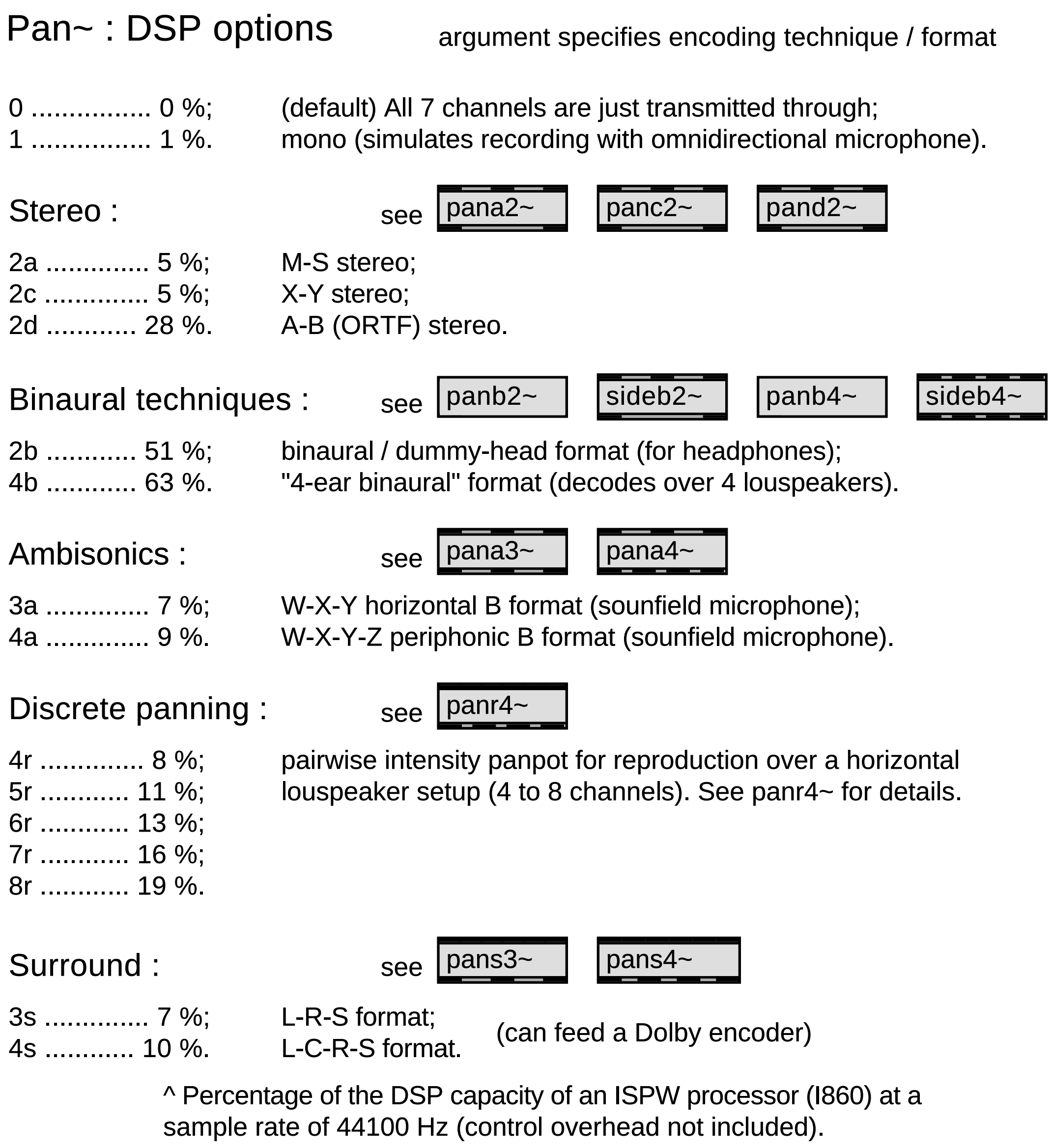
(images courtesy IRCAM)

Pan~: Signal Processing and low-level control This module receives seven signals: one *center* channel, two *sides* channels, and four *surround* channels (containing respectively the direct sound, the early reflections and the diffuse reverberation). **Pan~** can be configured to deliver 2 to 8 signals for feeding the loudspeaker system, and allows dynamic control of the apparent source localization with respect to the listener. The control interface in the **Pan_** object is divided in two sections: **source localization** and **loudspeaker system** configuration. The source localization is described in polar coordinates, using two control parameters:

- the apparent source azimuth (angle measured in the horizontal plane),
- the apparent source distance from the reference listening position.

Modifying the azimuth affects the distribution of the intensity of the *center* channel (direct sound) among the loudspeakers. The method used is derived from Chowning's algorithm [1, 7]. The distribution of the *surround* channels (containing the diffuse reverberation) is not affected by the source localization control. However, **Pan~** extends Chowning's method by allowing for the two *sides* channels (containing the early reflections) to rotate along with the center channel, according to the azimuth control.

Recently, **Pan~** has been extended to provide additional DSP options, such as:



Pan~ : DSP options

argument specifies encoding technique / format

0 .... 0 %; (default) All 7 channels are just transmitted through;
1 .... 1 %. mono (simulates recording with omnidirectional microphone).

Stereo : see pana2~ panc2~ pand2~

2a .... 5 %; M-S stereo;
2c .... 5 %; X-Y stereo;
2d .... 28 %. A-B (ORTF) stereo.

Binaural techniques : see panb2~ sideb2~ panb4~ sideb4~

2b .... 51 %; binaural / dummy-head format (for headphones);
4b .... 63 %. "4-ear binaural" format (decodes over 4 louspeakers).

Ambisonics : see pana3~ pana4~

3a .... 7 %; W-X-Y horizontal B format (sounfield microphone);
4a .... 9 %. W-X-Y-Z periphonic B format (sounfield microphone).

Discrete panning : see panr4~

4r .... 8 %; pairwise intensity panpot for reproduction over a horizontal
5r .... 11 %; louspeaker setup (4 to 8 channels). See panr4~ for details.
6r .... 13 %;
7r .... 16 %;
8r .... 19 %.

Surround : see pans3~ pans4~

3s .... 7 %; L-R-S format;
4s .... 10 %. L-C-R-S format.
(can feed a Dolby encoder)

^ Percentage of the DSP capacity of an ISPW processor (I860) at a sample rate of 44100 Hz (control overhead not included).

out~: This module can be used to apply spectral and temporal corrections to the output signals of the **Pan~** module, before sending these signals to the loudspeakers. Each channel undergoes an adjustable time delay and a double shelving filter of the type described in section 3.1.3. The filters can be used to equalize the frequency response of each loudspeaker separately. The delays can be used to make the signal propagation delays of all channels identical.