Jérôme Daniel
France Télécom R&D
DIH/IPS/ISI
Technopole Anticipa
2, Avenue Pierre Marzin
22307 Lannion Cedex
France
Phone: (+33) 2 96 05 27 96
Fax: (+33) 2 96 05 35 30
jerome.daniel@rd.francetelecom.fr
3D audio related web pages: http://gyronymo.free.fr

Position paper for the Campfire on Acoustic Rendering for Virtual Environments

Introduction: paper overview and additional resources

The area of expertise presented below in the first few sections issues substantially from my PhD thesis work, prepared at the Rennes Labs of France-Telecom R&D and recently defended (September 2000). It deals mainly with the reproduction techniques and the sound field representation that they are associated to, with the aim being to apply them to the 3D browsing in virtual environments. Among them, the ambisonic approach is more specifically developed: almost all of the aspects of the traditional first order systems are generalized to any higher order, for horizontal and full 3D reproduction configurations, and the usually referred psychoacoustic theories, based on the velocity and energy vectors, are thoroughly justified and interpreted.

For further information, the thesis document and the defense presentation (in french) are downloadable on my web pages (http://gyronymo.free.fr/audio3D/download_Thesis_PwPt.html) with english comments on each chapter, and an additional page (in english) gives commented sound and visual illustrations of higher order ambisonic rendering (see and hear: http://gyronymo.free.fr/audio3D/the_experimenter_corner.html). French and english abstracts are also available *via* http://gyronymo.free.fr/audio3D/accueil.html#lecture_audio3D.

The first section (3 pages) describes the ambisonic approach characteristics, the recent progress toward higher orders, and the expectations regarding its future.

The second section (3 other pages) opens a more general discussion on the reproduction techniques. The main classes of sound imaging strategies over loudspeakers (Amplitude Panning and Ambisonics, Transaural and Extended Transaural, WFS or Holophony) can be compared on the basis of acoustical considerations about the synthesized sound field. As a function of the chosen strategy and for given loudspeaker configurations, different compromises are achieved regarding the listening constraints, the satisfaction of natural localization mechanisms, the sound image accuracy, and the preservation of spatial qualities.

The third and last section (last page) briefly exposes current interests related to my recent activity in the 3D sound team of France Telecom R&D. Whereas the previous sections handle the reproduction, this last one deals with the content creation of virtual sound environments in a large sense, including the modeling of acoustical interactions (room effect, obstruction, etc…).

For these first two sections, the reproduction techniques are considered for their ability to reproduce the effect of each elementary event (wave front) of a pre-composed sound field, and in the end, to reproduce its macroscopic effect ("how preserved the global spatial qualities can be expected to be?").

First and higher order ambisonics

Brief overview

Ambisonics is worth being considered as a sound field *representation*, as a *sound imaging technique*, and as a whole reproduction *system*.

The ambisonic approach is based on *spherical harmonic decomposition of the acoustic field*, centered on the listener viewpoint. It has been known for a long time as a first order restricted form, which processes a minimal, **directional** sound field **encoding** through four components (B-format): W (pressure) and X, Y, Z (pressure gradient), offering easy sound field manipulations, such as rotations (see figures at http://gyronymo.free.fr/audio3D/accueil.html#choixsujet_audio3D). Ambisonic field can be encoded either acoustically, using a dedicated microphone, or synthetically, as a function of the directions of virtual sources and their associated reflections.

A **decoder** can be defined for various panoramic (2D) or periphonic (3D) loudspeaker rigs: it consists in matrixing ambisonic channels to feed the loudspeakers, in order to reproduce the original sound field at the listener place, or at least its perceptive effect. *Three primitive decoding solutions* had been defined for the first order systems to optimize the directional rendering in terms of the listening conditions: the LF-optimized (referred to as *"basic"*, later) and HF-optimized (*"max r_E"*) solutions, given by M.A.Gerzon for an ideal, centered listening, and *"in-phase"* decoding proposed by D.G.Malham for a collective, off-centered listening. Rendering can extend to headphones or a pair of loudspeakers *via* binaural techniques (virtual loudspeakers).

By considering in addition **higher order spherical harmonic components**, *the directional resolution* of the encoded sound scene *increases*. Quantitatively, the extended B-format consists of $K=(M+1)^2$ channels for a full 3D, M^{th} order representation, or only $K=2M+1$ channels for an horizontal restricted representation. The rendering requires more loudspeakers than ambisonic channels.

As a sound field representation based technique, Ambisonics is thus characterized by a **very appreciable versatility:**

- "Variable geometry" rendering (various loudspeaker configurations, plus possible headphone presentation)
- Ability to sound field transformations (rotations and perspectives deformations)
- "Variable resolution" sound field representation (scalability) used as a function of the transmission or/and the rendering capabilities
- "Variable listening area" decoding adaptability

As a system, Ambisonics has a quite **simple and low-cost implementation**, and offers **processing conveniences**:

All steps of the system are *simple linear operations* (substantially *matrix* operations, excepted the decoding for a binaural presentation), which are applied to the input or intermediary signals. These are: directional encoding of the sound field; optional sound field manipulations; optional mix of natural or synthetic sound fields; decoding (with optionally a low-cost *shelf-filtering*).

Note that the *decoding cost* doesn't depend on the original sound scene complexity (number of sources, reflections, etc.).

For a binaural presentation, decoding involves typically as many transfer functions as ambisonic channels. When dealing with many virtual sources, it can be interesting to use Ambisonics as an intermediate compact representation, in order to *factorize* the positional processing and to save CPU.

(Note: some emerging techniques dedicated to binaural synthesis do that with a better efficiency). Head-tracking can be handled by simply rotating the whole ambisonic field just before decoding.

As a rendering technique over loudspeakers, Ambisonics ensures **good predictability and homogeneity** of the rendered spatial qualities.

The encoding and decoding of each sound source (or phantom image) is equivalent to an *amplitude pan-pot*, thus the localization effect at the centered position can be predicted by the *velocity and energy vectors* ***V*** and ***E*** (ref thesis or any ambisonic related document). Since the decoder ensures that these vectors are compliant with the expected direction, *the directional information is preserved* (or controlled with virtual sources).

A *homogeneous* rendering is provided along all the directions; while ensuring the *loudspeaker "dematerialization"* (by avoiding to perceive them as individual sources). It also satisfies the naturalness of dynamic localization mechanisms (ITD and ILD variations due to head rotations, especially in low frequencies).

1st order system limitations: compared with the original "real" sound field experience, first order ambisonic rendering suffers from a lack of lateralization, which is felt as an elevation effect or as a *loss of image precision*. From a macroscopic point of view, considering a complex, reverberant field, the lack of lateral separation *may be* perceived as a partial loss of Spatial Impressions (S.I.) and envelopment (accompanied with a coloration effect).

Using higher order harmonics, which needs also more loudspeakers, allows to better benefit from the number of loudspeakers and their angular density (i.e. to use them more selectively). That way, the sound image robustness, its precision and the listening area are increased, and the spatial qualities better preserved thanks to a better lateral separation.

Recent progress: theoretical developments and understanding

Previous studies (Bamford95, Poletti96) have opened the way to the extension of ambisonic rendering to higher order, though offering partial view and extension of the approach. These have been completed by further studies (Daniel98, Nicol99, Furse&Malham99, Daniel00, etc.). In the following, I present the contributions issuing from my thesis work.

Technical and mathematical aspects [Ref chapter 3 of the thesis, plus defense presentation]

Most aspects of the traditional first order systems have been formally generalized to any higher order (for both 2D and 3D systems): the encoding, the decoding (major part of the work), and more partially the sound field transformations (rotations) and higher order microphone design.

For the **generic solving of the decoding problem**, underlying mathematics have been elucidated, in particular the directional sampling of the spherical harmonic basis (related to loudspeaker directions). Its regularity properties imply that the decoding matrix has a simple form, and that the local and global propagation properties (***V*** and ***E***) of the truncated sound field decomposition are preserved at the rendering. These concepts are also used for the design of higher order ambisonic microphones.

The primitive decoding solutions previously mentioned are generalized to higher orders into three families. They can be used separately or juxtaposed (per frequency band) to define an optimal decoder:

- The *"basic"* one optimizes the local centered reconstruction of the wave field (*i.e.* its extent regarding the wave length). It has to be used on a low frequency band, which narrows as the listening area extends.
- The *"max* r_E*"* one optimizes the "global propagation" ("global energy flow" ***E***), typically by "concentrating" the loudspeaker energy in the direction of the virtual source. It has to be used on the high frequency complementary band.
- The *"in-phase"* one minimizes directional artifacts and fluctuations when the listening area extends up to the loudspeaker perimeter.

Rendering prediction and characterization: "psychoacoustic" localization theories

Velocity and energy vectors (***V*** and ***E***, defined as the mean of the loudspeaker directions weighted by respectively the amplitude or the energy of their feedings) have been introduced by Gerzon (also referring to Makita) as representing the low and high frequency localization effect, and used as "psychoacoustic criteria" for the decoder optimization. It appeared necessary to clarify the foundations of these theories, in order to better characterize and interpret the expected spatial effect from these vectors.

For this purpose, ***V*** and ***E*** are first defined as characterizing respectively the local and the "global" sound propagation, then prediction laws of interaural difference are shown and their perceptive implications are interpreted as a function of head motions [sections 1.5, 2.2, 2.4 of the thesis]. The macroscopic interpretation (Spatial Impressions with a complex field) is also discussed.

An intrinsic link is shown between ambisonic representation (and its order *M*) and the potential properties of the rendered field (local reconstruction extent and "quality" of the global propagation), and as a consequence, the potential perceived spatial qualities (localization accuracy, image robustness, spatial impressions…).

Objective evaluations [Chapter 4] of localization cues (Spectra, ITD, ILD) issuing from the rendering confirm the contribution of higher orders and are correlated with the velocity and energy vector predictions. They are now supported by some additional sound demos (though with rather unrealistic examples: http://gyronymo.free.fr/audio3D/the_experimenter_corner.html).

Formal listening validations would have to be carried out. Moreover, generalized systems are still young or even not completely implemented. Their uses in interactive applications (within a complete spatialization environment, including room effect synthesis) still have to be more extensively experienced too.

The next future of higher order ambisonics

Extended ambisonic formats have certainly a future, but fast no past yet… How will they be used and found to be useful? The question involves many aspects.

- *A versatile use*: music or ambient sound recording; transmitting a room or space effect through *3D Impulse Responses*, mixing different sound scenes and factorizing positional processing, even for binaural presentation…
- *Rendering* high order ambisonics requires quite *a lot of loudspeakers*… as other rendering techniques like *Wave Field Synthesis* (see later) do. Thus adapted loudspeaker configurations are not a dream.
- *Implementation of extended B-format as an extension to the WAV-format* is being discussed.
- *A common destiny* of extended B-format: shared by Ambisonics and the binaural B-format strategy (Ref Jot, Larcher…)!
- *Sound field pickup*: higher order ambisonic microphones are in study. Their issue can be expected as a great step for the usefulness of ambisonic approach.
- There's a pool of ambisonics' defenders, still ready to promote such developments.

Opening a discussion: Ambisonics among other sound imaging techniques

The following discussion is based only on acoustical considerations about the synthesized sound field (and their perceptive implications), without worrying about system aspects like the transmission or the computation costs. The purpose is to highlight the potential of each strategy in terms of the sound image accuracy, the spatial quality preservation, the listening constraints and the satisfaction of natural localization mechanisms, all this, as a function of the number of loudspeakers involved. Sometimes, paradoxes will appear between the aim at satisfying natural hearing mechanisms, and the listening constraints. (Ref thesis + PowerPoint presentation: slide "principes de création d'image sonore" and the following ones).

Main classes of sound imaging strategies over loudspeakers

One can distinguish between at least three main classes of sound imaging strategies over loudspeakers:

- Using **amplitude differences** between loudspeaker signals (for each sound image), the loudspeakers being placed at the same distance from the center: *pair-wise pan-pot* (or reproduction issuing from *MS or XY stereophonic recordings*) and *Ambisonics*. Thus, **the contributing waves converge synchronously** at the center (thus **one focused point**), resulting (without the listener diffraction) in a **local, synthetic wave front** that has **uniform propagation properties** (apparent local direction and speed, characterized by the velocity vector) over the full frequency band (or over the bandwidths where amplitude ratio are constants), **extending** from the center **in proportion to the wavelength**.
- **Focusing on the field reconstruction at both ears** (thus **two focused points**, with account to the head diffraction): *Transaural or Stereo-Dipole, Double and Extended Transaural*.
- *Holophony (acoustic equivalent to Holography) /Wave Field Synthesis (WFS)*: **reconstructing the wave field over an area from its value on the area boundary** (Kirschhof Integral). Involving in practice a "sampled boundary", *i.e.* a **finite, discrete microphone/loudspeaker array**, reconstruction is quite **homogeneous over the whole area** for each frequency, but **spatial aliasing** occurs in a high frequency domain as a function of the spacing between loudspeakers.

A fourth class is omitted here – phantom source imaging using **time differences between loudspeakers** (issuing from *spaced microphones techniques*) – because it provides quite unpredictable (and wandering) sound images, though a better lateral decorrelation and enhanced spatial impressions, compared with reproduction issuing from coincident microphone techniques. (Note that it could be considered as a *very* degenerated case of holophonic methods.)

In the following, **we don't consider adaptive systems** (like head tracking cross-talk cancellation**).**

Comparison of systems will be made firstly with a *limited number of loudspeakers* (two speaker pan-pot, low order ambisonics, *versus* transaural and extended transaural) and a *single listener*, and secondly with *many loudspeakers* (high order ambisonics *versus* WFS/Holophony), with an *extended listening area or moving listeners*.

Preliminary: Some very general and evident laws

For the rendering of each elementary wave front, *interference figures*, which can be observed in the frequency domain, are created by combination of the contributing waves coming from loudspeakers.

"In all cases, the interference figures have a size or a spatial periodicity that is typically ***wavelength proportional****"*. This means that listening cues control becomes less stable or achievable as one considers a higher frequency domain, whereas things are quite easy with low frequencies, *i.e.* with wavelengths that are long enough regarding the listener scale. By the way, **all rendering techniques process similarly for (very) low frequencies**, and for a given loudspeaker configuration.

"It's as much difficult to reproduce the effect of a wave front (or a sound source), as its direction (or location) is far from the real, contributing sound sources (loudspeakers)".

- *"Difficult"* means "hard to achieve with stability and accuracy, or on a large area, or on a large frequency band". More technically, it needs more energy and implies the simultaneous participation of antagonist loudspeakers (thus a highly variant interference figure).
- *"The effect"* is in the end the perceptive effect, regarding static and dynamic listening mechanisms (localization cues and their variations by head rotation), or from an acoustic point of view, the sound field in the neighborhood of ears.

"The number of rendering control degrees is limited by the number of loudspeakers." The control degrees (or parameters) are typically the focused points (*e.g.* the ears, or the center) for the sound field reconstruction, and the axis along which variations are considered.

Limited number of loudspeakers, individual, centered listening

(Amplitude Pan-pot and Low Order Ambisonics *versus* Extended Transaural)

With only two frontal loudspeakers (traditional stereo *versus* transaural or stereo-dipole)

What is lacking: *traditional stereo* can control the direction (only frontal) of a synthetic wave front, but not its apparent propagation speed (not the natural sound celerity), while cross-talk cancellation is achieved only for given ear positions with *transaural*. As a consequence, variations of localization cues (especially ITD) by *slight head rotation* cannot be natural. This can be perceptively interpreted as: either an elevation effect ("under-lateralization") for images between loudspeakers; or a directional move ("over-lateralization") for images outside the loudspeaker span (only with transaural).

Sound scene extent and image accuracy: because of the cross-talk, traditional stereo offers only a smeared localization effect (predicted by the energy vector ***E*** in HF), especially for central images, and confines the sound scene within the frontal loudspeaker interval; transaural offers theoretically a full 3D sound scene with strong phantom images, but back-front reversals occur, probably because of contradictory cues variations by head rotation.

What are the freedom degrees: in both cases: moves are not critical in the median plane of the loudspeakers, including the front-back axis.

Image stability is critical with lateral head movements, depending on the lateral extent of the interference figure and its variance (amplitude). This lateral extent increases as the frequency decreases and as the loudspeakers narrow.

Compromise regarding the loudspeaker angle: in traditional stereo, loudspeakers are placed at +-30° as a compromise between a "not too confined" sound scene and a "not too poor" central imaging; applying the transaural approach to a +-5° speaker positioning (*stereo-dipole*) greatly enlarges the interference figure, thus the phantom image stability. [Footnote: Jerry Bauck "hybrid" system with two frontal pairs (substantially: transaural for low frequencies, stereo-dipole for higher frequencies).]

With four loudspeakers (1st order horizontal ambisonics *versus* double-transaural strategies)

What's improved: *With ambisonics*, sound scene extends to the full surround, while allowing synthetic wave fronts to have a natural propagation speed (thus correct dynamic lateralization in LF), but HF localization cues (ITD, ILD, spectral cues) still being smeared. *With double-transaural, i.e.* a transaural process distributed over a frontal and a back speaker pairs (ref Olivier Warusfel, IRCAM, or J-M Jot for binaural B-format rendering), back-front reversals do not longer occur, and it is even possible with slight refinement (proposed in my thesis) to provide natural ITD variations with slight (yaw) head rotations (at least with LF).

New constraints! Because both frontal and back loudspeaker pairs participate (especially for *lateral* virtual sources), an interference effect appears along the front-back axis, and the ear signal reconstruction is no longer stable considering front-back moves. A second positional constraint is added.

***Paradox and critical situation for the "double-transaural"*:** The "double transaural" and especially the "double-stereo-dipole" (speakers at +-5° and 180+-5°) are expected to be very

unfavorable to *lateral* virtual sources, forcing to a very strict ear positioning along the front-back axis with regard to small wavelengths (HF). This is in contradiction with the aim to allow slight head rotations and to satisfy dynamic localization mechanisms.

The problem stands in the fact that these are *minimal layouts regarding the number of parameters* to be controlled (here: four). There's *no such problem with Ambisonics* (centered focused point), for which *cross-talk is anyway involved* in sound image illusion and localization for any head orientation, though it smears HF cues and cannot provide images as precise and strong as Transaural ideally does.

[Note that the comparison could extend to a "minimal" 3D (*e.g.* cubic) configuration: a new positional constraint (along the vertical axis) is added in this case.]

Increasing the number of loudspeakers: This paradox is progressively removed when loudspeakers are added without increasing the number of control parameters (*i.e.* without adding new dimensions). Comparatively, with more loudspeakers and higher order ambisonics, HF cues are less and less smeared while always featuring a natural dynamic localization.

It is likely that both kinds of rendering would converge, but Ambisonics is much easier to implement than Extended Transaural.

Many loudspeakers for an extended area (High Order Ambisonics *versus* WFS)

A concise comparison is given is the following table.

Rendering properties	**High Order Ambisonics**	**Wave Field Synthesis**
Sound field reconstruction (as the order increases)	Radial expansion (*kr*), wavelength proportional	Spectral expansion (*f*), uniform over the area
Loc. characterization outside the reconstruction domain	Energy vector ***E*** (HF/off-centered)	No prediction (above the spatial aliasing frequency)
Reference viewpoint	Unique (centered listener) and extrapolated	Global
Sound image projection (converging point of perceived directions from all listening positions)	Over the loudspeaker array (like usual visual image projections) (see comment *)	Beyond the array, with respect to the original source distance (like holographic images)

(*) To be more exact, high order ambisonics is able to reproduce the effect of sound sources *beyond* the loudspeaker array, but *only within the reconstruction domain*: it only requires compensating the near field effect of the loudspeakers.

The last two lines of the table introduce *the question of the audio-visual coherency*, since a true holographic visual rendering is not achieved. Such a coherency seems to be better achieved with High Order Ambisonics, which tends to act as the usual, visual projection (with the image corresponding to *one* viewpoint). Despite of absolute directional distortions perceived at off-center positions, perspective information is preserved through the relation between direct and reverberated sound. However, the level distortion caused by loudspeaker proximity can be a problem and its effect should be further evaluated.

Conclusion: Systems have been compared on the basis of objective arguments. It would be worth confronting these expectations to practical, audible experiences!

Bibliography

Refer to the work of: Gerzon, Malham, Bamford, Poletti, Daniel, Nicol for Ambisonics; Larcher, Jot, Warusfel for binaural B-format and double-transaural; Nicol and Delft University for WFS/Holophony.

Current activity and special interests: acoustic modeling and content creation tools

As a complement to the question of the sound field reproduction (or sound imaging) treated just above, my current interests are rather concerned with:

1. The content production of virtual sound environments;
2. The efficient integration of advanced technologies using existing hardware.

The first point, beyond the *ergonomics* of Human-Computer Interfaces of content creation tools, involves several aspects: *refinement of virtual acoustic modeling* for a more immersing and interactive rendering (room effect and coupling, occlusion and obstruction); its *translation* into parameters of standardized description formats; the extension of *description formats*.

The second purpose deals with technical questions such as the description formats, the plat-form variability, and the repartition of processing tasks between hardware and software. Regarding current API features, an additional question is *the control or the choice of the sound imaging technique* at the final stage of the rendering (which doesn't seem to be proposed yet).

It is hoped that the emphasized aspects will be further discussed during the Campfire.