3D AUDIO AS AN INFORMATION ENVIRONMENT

PETER P.LENNOX[1] JOHN M.VAUGHAN MSC[2] AND DR TONY MYATT[3]

[1]Research Student, Department of Music, University of York, York YO10 5DD, United Kingdom
ppl100@york.ac.uk
[2]Research Associate, Department of Music, University of York, York YO10 5DD, United Kingdom
jmv100@york.ac.uk
[3]Director of Music Technology Research, Department of Music, University of York, York YO10 5DD, United Kingdom
tone@cage.york.ac.uk

To progress from surround sound to true 3-D requires a more psychoacoustically complete depiction of the audio environment. This paper defines an important element of the interaction of sounding objects and their environment that we have termed 'ambience labelling information'. However, complex scenes can be cartoonified in order to preserve only perceptually significant information and thus greatly simplify rendering.

INTRODUCTION

This paper is the second in a series that describes a model of human auditory spatial perception based upon contemporary thinking in psychology, psychophysics and neurobiology, informing in particular the production and reproduction of 3D audio. It leads to a view of human audition that is quite different from the theories that underlie the use and design of most current surround sound technologies.

We have elsewhere[1] proposed the term 'Perceptual Space' in order to emphasise the crucial distinction between abstract physical (Euclidean) space and that perceived by human listeners. The key features of perceptual space are the concepts of 'perceptual significance' and 'ambience labelling information'. Perceptual significance has been discussed in a previous paper[1], and highlights the finding that spatial perception is shaped by the adaptive need to rapidly codify features of the environment in accordance with relevance hierarchies, elucidating and interpreting behaviour partly through selective inattention to the background.

Ambience labelling information concerns the relationship between sounding objects and their environment. The term refers to a class of information available in real environments, apprehended using dedicated preconscious processes. The significance of this type of information can be exemplified by the fact that, from the point of view of human auditory spatial perception, the sound of an object moving is determined to a much greater extent by the change over time of the object's audio relationship with its environment and other sound objects, than by the change over time of subtended angle to the listener, and indeed the character of movement itself is apprehensible quite independently of changes in interaural differences of any sort. Surround sound systems that rely simply on mapping geometric relationships between sound objects and the listener, however accurately, will ultimately fail to produce believable audio environments, as the crucial information channels required for the cognitive understanding of the environment are missing. It is this factor which, we believe, characterises the distinction between surround sound and what we term True-3D, or, for small-scale listening circumstances, soundfields and sound environments.

However, the acoustic relationships of sounding objects with their environment are well understood, and from a technical point of view facilities exist for describing these in audio terms. The difficulty is that accurately describing the interaction of multiple, perhaps moving, sound sources in a feature-rich environment can become fantastically complex and require immense processing power. Again, the principles of perceptual significance inform this process, because it turns out that it is not actually necessary to describe in detail the interaction of sound sources, provided that the perceptually important information is delivered. This highlights the result that there are certain relevant classes of information to which the ear-brain is adaptively predisposed. The process of deriving simplified representations of complex real-world situations that still satisfy the most important information channels to the ear-brain we have termed 'cartoonification'.

The paper will provide a comprehensive description of ambience labelling information and how it relates to surround sound production and reproduction. It will also demonstrate why a description, in audio terms, of the interaction of sound sources with their environment is essential to the synthesis of believable soundfields. We will then go on to show how the principles of perceptual significance can be used in order to cartoonify complex changing audio scenes, extracting key information that greatly simplifies rendering. Audio production in surround does require a fundamental reappraisal of sound design, and while many engineers understand the inadequacy of conventional approaches, new techniques, and in particular technologies that are only now becoming available, are alien and unfamiliar. Crucially, our work demonstrates that relatively much greater attention needs to be paid to the space, perceptual and physical, in which sounding objects find themselves, in order to achieve something that is more than just wraparound stereo.

1. AMBIENCE LABELLING INFORMATION

1.1 Context and Envelopment

The challenge in soundfield synthesis is about 'externalisation' - what makes the world and its things 'out there'. It is precisely this 'out-thereness' that is generally missing in many artificial soundfields, is present in many ambisonic recordings, and is, we believe, the main difference between a soundfield and a sound environment. We suggest that there is a class of information that is perceptually important but which we don't 'focus' on: 'background'. Without this background, the 'foreground' objects of perception don't actually make sense, and we might regard this background as a *context* for sounding objects, helping us to discern and position them.

In sound-environments, 'ambience labelling information' *is* the context, in that it *is* the sound of the environment in which objects find themselves. So, in a given sound environment, we don't just hear the sounding *objects* in that environment. Indeed, often the majority of the sound in an environment is reflected sound, but a rich *pattern* of reflected sound, which, being inhomogeneous, serves to anchor the objects 'out there', irrespective of specific percipient positions.

Existing surround-sound systems conform to what we understand as 'soundfield production systems', attempting to accurately produce the salient plane-wave directional characteristics of sounds in a given soundfield, for a specific listener position. The intention is to generate specific, physiologically measurable sensory events, leading to specific auditory spatial perceptions equivalent to those that would pertain to a specified 'sound environment'. The hope is that the 'information' about, caused by, or in the specified sound environment can be portrayed, for a defined listener position, in the artificial soundfield.

We use the term 'ambience labelling information' to describe this specific, physical property of auditory environments. We acknowledge that the use of the term 'information' in this context is problematic, and that some clarification is required.

1.2 Some Facts on Information

For example, in Gibson's view:

"Information .... does not consist of signals to be interpreted, nor data that must be supplemented from a store house of knowledge. I am suggesting nothing less than the hypothesis that meanings are not subjective contributions but objective facts. I prefer to call the meanings of things their affordances, that is, what they afford the observer. The meanings of things in this sense are perceptible properties of things...."[2]

In fact, Gibson went as far as to say that these meanings were apprehensible through direct unconscious processes in a way that did not involve cognition at all[3]. In the alternative view, and as a direct response to the above quote, Albert Bregman states:

"I have trouble with the idea that 'information' or 'meaning' is in things. Let me focus my discussion on 'information', although an analogous argument could be made for meaning.

"Using the word 'information' to refer to any physical fact or relation in the world broadens the idea of information too much. The thing that is in the world, is 'pattern', not 'information'. Pattern only becomes information when it is used in a communication system to send a message from a sender to a receiver. Furthermore, it is important to recognise that the kinds of pattern that count as information will depend on the properties of the receiver. If you send Morse code to a TV receiver, the signal isn't information, but noise.

"The existence of 'information' is a property of the whole system, not the patterns taken alone. Since the properties of the brain (or *any* receiver) figure in the definition of 'information', then information

can't exist, by itself, in the environment. However, 'pattern' can."[4]

In other words, in this classical approach 'meaning' of 'information' is, with respect to the physical, spatial world, either epiphenomenal (having no direct causal relationship), or metaphenomenal (an emergent property of the environment-percipient relationship). However, a parsimonious definition of information does seem to leave something unsaid about the nature of such patterns - they become vague and unquantifiable, devoid of 'meaning'.

In any event, we assume that the purpose of a surround sound system is to engender quite specific spatial information. The interesting question here is whether a sound environment *can* be expressed in bare spatial terms as 'patterns that constitute a soundfield', with individual percipients contributing subjective impressions of meaning, or not. We would suggest not, of course, and that the difference between soundfields and sound environments is qualitative. Nevertheless, our argument that 'surround sound' is an attempt to depict a sound environment, and that such environments may be accessible rather more directly than through the painstaking application of soundfield technologies, leads us to propose that there should be certain 'patterns' or 'information-classes' in the environment, not inherently subjective, that are therefore subject to measurement. These we term, with apologies to our detractors, 'information'.

1.3 The Texture of Space

Of course, Gibson was concerned primarily with visual perception, and we are cautious of metaphorical comparisons between the visual and auditory worlds - indeed this paper proposes that auditory spatial perception and visual spatial perception may be qualitatively different, and the extended use of visual models may be detrimental to the development of understanding of the true nature of 'auditory space'.

These reservations notwithstanding, Gibson's concept of optical 'texture' as a primary cue to 'location perception' (in particular 'distance-from-me' or 'size-of' assessments) may well have an auditory parallel: what we suggest is that the auditory environment is also 'textured' and that we have the neurological structure to process for this texture directly. This texture is what we term 'ambience labelling information'. Our experiments lead us to the conclusion that 'distance-from-me' judgements are every bit as good in the auditory domain as in the visual (and of course better for objects hidden by other occluding features). We suspect that the origin of this texture lies in what happens to transients when they collide with features of 'real' environments: with each reiteration-and-subsequent recombination with unreflected sound, they become temporally blurred and 'compressed', resulting in perceptually noticeable spectral changes.

In a previous paper[1] we stated that perhaps the most significant insight that modern psychoacoustics has to offer the development of 3-D audio is the realisation that audio spatial perception is primarily a time-domain process. In that paper we went on to say, in drawing the distinction between visual and auditory perception, that:

"The key to individuating objects is our ability to discern 'edges' - the discontinuity between the optical characteristics of adjacent surfaces. This ability allows us to recognise 'shape'.

"By contrast, in the auditory domain, we have no need, or mechanism, to delineate 'edges' in order to detect organisms. We make quite different discernments about objects in the auditory world, whose perceptual edges are temporal rather than spatial."[1]

Because so much of the research into localisation acuity is performed in conditions that are (by necessity in order to factor out varieties of influences) unnatural, it may not provide a particularly good view of localisation processes in real environments. Certainly, the commonly-held view that localisation is rather poor at high frequencies seems not to hold with our experience that ambience labelling information is very effectively conveyed in the high frequencies. Temporal differences at the ears are only system-induced using conventional techniques at low frequencies, which is informed by an outmoded view that the temporal differences associated with high frequency signals could not be represented neurologically. The more complete view tells us quite the reverse: it is in fact the temporal relationships between self-similar audio components at *high frequencies* that primarily give rise to good individuation and localisation information. This is undoubtedly due to the importance of transients in allowing the characteristics of the audio space to be determined, and it is certainly true that transients offer perhaps the least ambiguous information to the auto- and cross-correlation processes that are known to

operate[e.g.5]. Transients are clearly crucial in establishing precedence, and although we do not yet have the experimental data to confirm this, it may be that this is a further manifestation of a common predisposition to the apprehension of novelty[6][7].

1.4 What and Where

We have elsewhere[1] highlighted the considerable neurobiological evidence that neurological representation is provided of multiple ‘what’ and ‘where’ processing streams in the visual system[e.g.8] along with the suggestion that analogous processes may exist in the auditory domain. The importance of multiple parallel processing 'systems' lies in the evolutionary advantage conferred: information about 'things' may be perceived in drastically sub-optimal conditions, provided at least some cues remain. Further, particular neuronal populations can, by specialising in particular perceptual features, prove more competent in their response times; a large complex task can functionally be decomposed into smaller, more manageable ones.

We stated that:

"…the ‘what’ representation yielded in the audio mode is different from, complementary to, and as richly detailed as that derived from the visual mode. Whilst form is not as well rendered, material composition, structure, type-of ‘what’ and size-of ‘what’ often are. On first consideration, our ‘where’ representation seems less detailed in the auditory mode, but this is to entertain the outdated misconception of sensation being equivalent to perception and thinking of senses as competitive/hierarchical votive systems. In the complementary and overlapping model, audition is as likely to occasion foveation as is vision. Furthermore, our audition is able to render ‘where’ information unavailable to vision. For visually occluded objects, either behind the perceiver or in another, presumably adjacent, space, information is readily available to audio perception as to rate of movement, change of movement and even reason for movement. This type of information may often be significant in determining call to action. It is available completely independently of changes in inter aural differences. We have termed this type of ‘where’ information ambience labelling information."[1]

Although we do not feel that the delineation is very strong and there is considerable overlap, we have characterised perceptual significance as being concerned with 'whats' and ambience labelling as providing information about 'wheres'. This is of course an oversimplification, because what we also contend is that perceptual significance is crucially concerned with the apprehension of behavioural 'affordances' (to use Gibson's term) that occur in priority to and independently of the construction of 'what' and 'where' perceptions. Nevertheless, it is useful to explore what the nature of this 'where' information might be, because it characterises well what type of information is presented and what adaptive significance is attached to it.

We do contend that ambience labelling information is 'objective', that is, it is extrinsic to any particular percipient and should therefore be measurable. This is the reason why we do not feel constrained by arguments surrounding the meaning and nature of 'information'. Regardless of where the information content arises, whether it is in the interaction of sounding objects and their environment, or whether it is truly an emergent property of the processes by which such patterns and modulation are apprehended is largely irrelevant, because it is clear that the mapping of these physical characteristics of the audio environment onto spatial perceptions is direct and coherent.

The ambience label provides information about the features and surfaces around and in the vicinity of a sounding object, thus being well-represented in environments rich in features and being correspondingly scarce in those relatively bare of features (and of course entirely absent in anechoic conditions). It may be useful for locating objects relative to surrounding features, but of course of itself is not 'distance information' which is inherently tied to a specific percipient-position. Although good specific comparative evidence from the literature is rather scarce on this point, it is our firm contention that ambience labelling information, in most real environments, directly contributes to localisation[9], so that anechoic localisation, based on interaural/pinna cues alone, is inferior. We believe this to be a crucial point, because the interaural approach to localisation so often assumes that the signals present in real environments in addition to the 'direct' signal can only serve to confuse the ear-brain. Moreover, distance perception is quite strongly assisted by the availability of ambience labelling information, which is a less contentious point[e.g.10]. Certainly, interaural theory is particularly weak in explaining distance perception, but it is clear that the perceptual apprehension of proximity, and in particular change-of-proximity (i.e. 'comingness' or 'looming') is of crucial adaptive importance, and indeed strong evidence is emerging of an adaptive bias towards the perception of comingness[11].

2. CARTOONIFICATION

2.1 'Meaning'-Based Compression

Of course, the relationships of sound objects with their physical environment is actually very well understood from an acoustical point-of-view, and from a technical point-of-view signal processing techniques exist for describing these in audio terms. The difficulty is that, in accurately describing the interaction of multiple, perhaps moving, sound sources in a feature-rich environment, achieving coherence of a very large numbers of parameters may require immense processing power. Of course, audio engineers have long been used to abstracting complexity in order to achieve sonically acceptably results efficiently. A 'large hall' reverb setting would be an example. However, the process of expanding these notions to surround or 3D audio adds layers of complexity that make the abstraction process itself ever more critical. It becomes necessary to extract and codify the features of the simulated audio space that provide perceptually the most significant cues, and ensure that processing power is devoted to rendering these to the required degree of accuracy. Although it turns out that the techniques for achieving believability in sound fields are often highly task-specific, we have derived a frame of reference that we believe will assist in the production of surround material that achieves true depth of significance. At the core of this is the need to attend in a much more specific and technical manner to the space within which the postulated or real sound objects purport to exist. Only then can the interaction of postulated or real sound objects with each other and their environment be modelled. In addition, careful attention should be paid to the highly perceptually significant elements of the behaviour of the sound objects. Once the perceptually important features of the sound environment have been defined, these can be selectively modelled to an extent which is much more complete than other parameters of the environment, yielding a result that is efficient in terms of signal processing yet perceptually satisfying. This is the process that we have termed 'cartoonification'.

It is really a commonsensical notion, drawing (inevitably) on visual metaphors. The skill in efficiently 'cartoonifying' lies not only in stripping away unnecessary information, but also in exaggerating particular information features to facilitate particular intended perceptions, unambiguously. This involves *distortion* of information to engender *accuracy* of perception. So a skilful cartoonist can, with a few strokes of a pen, depict a *particular* person, in a *particular* state of motion, with perhaps even some notion of what the person is *thinking* or *intending.* Obviously, this relies on a common subjective vernacular, and that this vernacular features *exaggeration* as a key component of fast perception.

The interesting question is whether auditory spatial perception relies on *cognitive* exaggeration of key features, and whether these features are common to the generality of human perception. We believe there is considerable evidence to support this view, in the acceptability of the use of artificial reverberation to convey 'distance' information, in the perception of 'stereo' sound fields, and where microphone recordings of moving objects need not feature changes in interaural differences to prove acceptable in specific instances.

In his presentation to the AES 16th International Conference on Spatial Sound Reproduction, discussing a 3D audio headphone display referred to in his paper[12], Durand Begault stated that:

".reverberation has been shown to dramatically increase the externalisation of stimuli relative to non-reverberated stimuli, in one case from 2% to 90% ... It may be possible to mitigate reversal errors by establishing a 'cognitive map' of the acoustical features of reverberant cues."

We would characterise this last point as an example of the cartoonification of ambience labelling information.

Inevitably, prior to proper quantification of 'perceptual cartoonification', no global approach to signal processing for cartoonification will suffice for the wide variety of tasks expected of 3D auditory display technology; individual manipulations must be task-specific, programme material-specific and display technology-specific. In a previous paper[1] we discussed the distinction between surround sound presentations and what we termed 'True-3D' audio. Partly this was aimed at determining assessment criteria for the depth of illusion of sound (re)production systems. Although it was not our contention that we need aspire to the creation of a soundfield that will inevitably be confused with the real world, we did assume certain necessary attributes of 'realism'. Believability was one. For concert situations at least, surround sound systems fail to realise the potential for explorability, either through movement or selective attention. We applied this as a 'minimum requirement' for 'True 3-D' systems.

However, we acknowledge that the criterion of physical explorability is unnecessary and irrelevant with respect to small-scale systems designed primarily for single-perceiver perspectives. We would still apply the criterion of believability for 3D audio-only presentations, which is wholly a product of the apperception of texture as it pertains to 'out-thereness', although the relevance even of this is less clear when the audio material is primarily in support of a visual presentation. In such circumstances, the authors would question the appropriateness of surround panning techniques that coherently localise audio objects outside the visual presentation field. This does not however negate the usefulness of ambience labelling techniques in such circumstances, because an understanding of the concepts allows deliberate blurring of directional information, for example, in order to facilitate a strong sense of 'over-thereness' without attendant and potentially distracting localisation cues.

In fact, to achieve technical realisation of any audio illusion, the concepts of ambience labelling are crucial: often the most significant bar to achieving audio illusions (or depth-of-illusion) is the ambience label of the loudspeakers used to generate them. An example of this is to be found in the way 'precedence effect' can interfere with imagery for a listener not in or near the 'sweet spot'. This is important because the only way to compensate for this would be to attempt to 'perceptually de-localise' the loudspeakers by inhibiting in some way their ambience label. This explains why the sonic quality, or depth of illusion, improves so dramatically through the use of distributed arrays of loudspeakers, such as in an ambisonic field, in preference to the minimum required.

To summarise this section, immersion in a believable soundfield that engenders a sense of 'out-thereness' requires a process of removing perceptually spurious, and exaggerating perceptually relevant ambience labelling information. To achieve a systematic approach to this we need to know what *generalisations* we can identify with regard to ambience labelling information in real environments, and what selection features can be said to be *general* to human perception. For instance, one aspect of believability has to do with the fact that 'real environments' have 'shape', as do real 'things'. These shapes are usually not symmetrical, and this useful generality of the world-about-us may underpin a great deal of our spatial perception.

2.2 Perceptual Space versus Three-Dimensional Space

We have used the term 'perceptual space' to refer to a 'space' that is not the classical physical space but an information environment. This arose from the appreciation that the world is not comprehended solely in Euclidean terms. The units of our 'perceptual' space are not absolute, objectively measurable, nor independent of a perceiver's viewpoint. They are instead relative units, their values varying according to the perceiver's assessment of the importance of various features of an environment.

In fact we hypothesise that the primitive 'building blocks' with which we understand the spaces we inhabit have little to do with Euclidean space, but consist primarily of what we might term 'units of urgency'. So for example, it is more 'urgent' to perceive that something potentially dangerous is 'coming' than it is to have a complete and accurate cognitive representation of the local topography. We feel that arguments to the effect that such a representation must be necessary *before* the 'comingness' of a 'thing' could possibly be perceived, are unsupported. Put simply, 'comingness' is perceptible before 'what' is coming from 'where': the latter are more sophisticated developments built on the former.

To hypothesise this would require that the information pertaining to *comingness* must be found in the environment, independently of topographical knowledge, and must be obtainable quite simply, without detailed information of *what* it is that is coming. We further propose that, in a relevance hierarchy built on survival-related urgency, the most important 'what' that *could come* is a predator, hence *'behaviour'* that may convey '*intention'* will require the most rapid processing. Behaviour detection of *organisms* must similarly rely on fairly simple *stimulus information in* the environment. Whilst a notion that this requires at least some representation of the 'space' surrounding both the perceiver and the potential predator seems reasonable, this does not require a detailed 'map' of all local features. There is evidence that such map-like cognitive representations might be most closely represented in phylogenetically higher orders of organisms, but we suppose that spaces relevant to human perceivers can often be quite simply categorised according to subjective needs, and that, for many tasks, more detailed models may in fact be unnecessarily cumbersome in use.

The primary perceptual spaces of interest are 'my space', or what we might term the 'immediate zone',

and, to a slightly lesser extent, any adjacent space; beyond those we have what we call 'distant'. However, it is important to bear in mind that the impact of adjacent spaces on the ambience labelling characteristics of proximate sounding objects can be significant. The cartoonification process must include, as a starting point, some notion of the major characteristics, principally the gross shape and nature of any potential occluding features, of these spaces.

2.3 The Shape of Perceptual Space

The grossest generalisation of the nature of an environment is the shape of 'my' space and that of any adjacent spaces, the features surrounding me, and the surfaces surrounding sounding objects. In section 2.1, we mentioned the need to define the generalities in real environments that might be conveyed in ambience labelling information, and be sufficient to instantiate any of several perceptual processes.

One of the most significant and consistent features of real environments is the ground surface. Extending under sounding objects and percipients alike it reflects *and* absorbs and even sometimes transmits 'direct' sound; any environmental feature *less* like a 'point source' is hard to imagine. Often the most proximate reflecting surface, and at a (generally) known distance from the percipient's ears, the ground surface is so ubiquitous as to be 'beneath notice', yet its effect on auditory spaces is pervasive. If our conjecture as to the importance of such a feature is correct, then the absence of this 'ground effect' should prove detrimental to the perception of 'distance-from-me'. Informal experiments with volunteers speaking and listening at a height of 6 metres, separated by a fairly non-reverberant space, seem to confirm our predictions that, without this ground effect, *perceptual* distances are lessened considerably, when compared to actual distances. This ground effect and its importance to perceptual significance certainly seems worthy of more controlled investigation; we speculate that this importance will be found to increase with proximity between sound source and percipient.

But the fact that this effect is so difficult to introspect on highlights another important aspect of 'perceptual space'; namely that some of the parallel processes that constitute perception contribute to *selective attention* by processing for selective *in*-attention. That is to say, certain regularities common to (most) environments should not be *noticed* in themselves, they are of low 'perceptual significance'. Only *unexpected* disturbance of these regularities should instantiate attentional processes; thus, non-coherence, or inconsistency in the ambience labelling fabric of an artificial soundfield may prove unduly disturbing.

The immediate zone is poorly depicted in commercially available surround sound systems, which generally cannot utilise the space within the speaker array. Depth-of-field for such systems starts at the perimeter of the array. The immediate zone is delimited by the immediate boundaries that are primarily giving rise to ambience labelling information, and thus to a sense of envelopment. In adjacent space, ambience labelling information is not so clearly rendered, and generally comes from a specific direction. The ambience label of sounding objects in an adjacent space primarily provides information about the adjacent space, and generally very little about the immediate space. Interaural difference information provides clues as to the route by which the sound reaches the perceiver, but the ambience labelling information will primarily be that of the originating space, provided that that adjacent spaces are not too strongly excited. Therefore, localisation of objects in the distant zone will be very imprecise, although apprehension of approach (in particular[11]) or retreat is more successful. The distant zone will generally be perceived through apertures in the boundaries of the immediate space. For sounding objects in distant spaces, judgements of 'facingness' (with respect to me) may be relatively poor. However, *change*-of-facingness with respect to features local to the sounding object may be quite good. In fact, facingness may sometimes be confused with 'distance'; objects facing away from me tend to sound relatively further away than objects facing me.

To summarise this section, we have said that 'space' can be loosely decomposed into 'my space', 'adjacent space(s)' and 'distant'. In sound terms, each of these can be further decomposed into 'things' (that instantiate sound) and 'place-features' (that do not 'make' sound). The *things* are the perceptual foreground items, to which we pay selective attention, whilst the place-features are *heard*, but not *attended* to. Considering the spaces purely in terms of place-features, for a moment, 'my space' should command greater perceptual processing because a) the things *in* it are potentially more urgent (according to *perceptual significance*, see 2.5) and b) there is more textural inhomogeneity available in the ambience labelling information.

2.4 The Parameters of Thingness

In the same way that we identify generalities common to most environments in the previous section, we clearly need a reliable, common physical property of *things* that may be readily discernible in a wide variety of circumstances. This is all that is required of an instantiating cue, as more complex and flexible cognitive processes can rapidly be deployed on recognition of the *possibility* of 'thingness'.

Fortunately, the problem is somewhat simpler for auditory perception than it is for visual perception; basically, if it makes a sound, it is probably a 'thing', and in interaural terms, *precedence effect* has been shown to be very powerful. But we have already described many aspects of 'spaces' where interaural difference information may not suffice. Our consideration of supplementary characteristics suggests two main useful candidates: 'behaviour' and 'body'. The things in the world that make sounds all exhibit movement (of some sort, even if stationary) and all have physical dimensions, filled with matter. Obviously, the important (by far) kind of 'thing' we generally need to identify quickly is 'organism'; one large enough to significantly affect our physical well-being is presumably particularly perceptually interesting. In many environments, the majority of sound-events we hear *are*, or are *caused* by, organic sources.

In a previous paper[1], we offered the term 'Perceptual Significance' to define the characteristics of an audio event that might predispose apprehension. In that paper, we talked of 'facingness' as being an acoustically coherent property, particularly of organisms, and its physical properties are reasonably well understood, measurable and reproducible: it is relatively easy to control the perceptually important features of facingness. Facingness arises from the directional inhomogeneity-of-sound-output of most sounding objects, and the modulations of patterns/information that occur through body occlusion are clearly a feature of ambience labelling. Fairly simple signal processing treatments for facingness are therefore possible, and while their use in conventional musical displays may be limited, we foresee considerable potential for the selection of material, through facingness, in auditory displays where multiple channels of auditory information are being presented.

Of course, the features that constitute aspects of ambience labelling information, such as facingness, are especially suited to describe the 'behaviour' of sounding objects in and through their local environment - this is not just a case of Doppler effect, as it includes the timbral changes due to comb-filter effects as the early reflection patterns change with movement. Behaviour, and the prediction of behaviour, is a crucial class of information that an ecological view makes us presume our perceptual systems are predisposed to apprehend, and therefore quite subtle and simple-to-achieve changes in the ambience label can be used in order to engender strong emotional or cognitive responses, such as fear or attraction.

This area is crucial: It is our contention that human perception has an overwhelming bias towards the apprehension of behaviour, and more generally the parameters of what we would term 'thingness'. However, this is a substantial area for discussion that we cannot hope to cover in this paper, and so it will form the subject of a future paper.

2.5 Perceptual Significance

The authors employ the term 'perceptual significance' to emphasise the ways in which cognitive functions select for attention those information-yielding properties that have the potential for facilitating the most useful predictions. We have already seen how our perceptual systems distort physical space and select for the apprehension and prediction of behaviour. In a previous paper[1] we postulated that the number of elements in the significance hierarchy was about seven, which in turn informs, for instance, the maximum number of different degrees of relevance that can be used in simultaneous multi-channel information presentations using auditory displays. It also determines the level of complexity of a synthesised audio environment at which individuation of its difference elements according to perceptual significance will cease. Of course, there is room for debate as to what extent the range of perceptual significance in an artificial soundfield can ever approximate that in a real environment, but there is a sense in which (for surround sound systems in small rooms) the physical depth of field implied by a system can be said to correspond somewhat to a psychological 'depth of significance'.

2.6 From Surround to 3D

We contend that the synthesis of spatially 'accurate' complex environments is both technologically unfeasible (in most practical situations) and perceptually unnecessary. The authors' experience has shown that the current approach to representing spaces in a cartoonified form, such as presets on reverb units, accords well with our appreciation of the perceptual insignificance of detailed spatial

mapping. There is no reason to suppose that similar cartoonified representations will not be equally successful in 3D presentations, although proximity may prove a useful additional parameter. Furthermore, we suggest that the *minimum r*equirement for believability would be a stable and consistent audible 'background' (context), against which sounding objects 'make sense'. The most important application for ambience labelling information is to achieve a tangible sense of envelopment. This cannot be achieved simply by surrounding a listener with sound sources - it is, in real environments, available when a single source at a particular location interacts with its environment (for example a single speaker in a room engenders a sense of envelopment). Envelopment is achieved by filling in the background/texture, by placing a sound source coherently within a space.

Perceptually, we have found that one of the principal criteria for achieving envelopment is for the acoustical components of the various audio interactions to be coherent and consistent, that is, that the acoustical representation of the environment remains constant in the presence of a variety of sounding objects, which means that, in a synthesised soundfield, they must respond in the same way in order to be believable. In fact, in modelling synthetic environments, complexity and accuracy of detail is largely unnecessary, provided that a sufficient degree of consistency is achieved. Ambience labelling also explains why it is often difficult to achieve believable representations of sound environments using loudspeakers using current technologies, and this is because of the ambience labelling characteristics of the loudspeakers themselves. Often, in a synthesised soundfield, the most stable and consistent part of ambience labelling information is occasioned by the loudspeakers themselves. It may also explain why ambisonic recordings are not so badly affected in this way, in that strongly coherent ambience labelling information is contained in the recording, but it is certainly true that the technology itself plays some part in this, because the ability of a system to convey detailed ambience labelling information without significant distortion is crucial, something which is clearly a feature of Ambisonics[13]. We might even conclude that the effect on reproduction of ambience labelling information is more crucial than that of sound object itself, especially if the latter is localised at or very near a speaker. A corollary of this, which we have briefly mentioned before, is that it becomes easier to render believable soundfields over distributed multi-speaker arrays. We can suppose that the reasons for this include the fact that a) the ambience labelling information of presented material more completed rendered; b) SPLs at each speaker are relatively lower so that its acoustic interaction with its environment is less localisable to that speaker; and c) there is a lessening of image distortions due to precedence effect for non-ideally positioned listeners. Furthermore, it has been shown that the action of the loudspeaker itself results in transients that are strongly perceptually significant (referred to in [14]), and a more diffuse presentation will certainly ameliorate this to some extent. However, what we are saying here is that more loudspeakers results in a potentially more believable envelopment, rather than simply better localisation of sounds. Significantly, we have found that rendering conventional surround sound (even panpot stereo) material into ambisonic format also gives a perceptually more involving result, as can small amounts of widely-spread reverb added to panpot stereo, which accords well with Begault's findings mentioned previously.

It appears, therefore, that sound reproduction systems are not all equally capable of conveying ambience labelling information. While Ambisonics appears to be particularly effective in this respect, 5-speaker systems are always going to place limitations on the believability of the result. There are, however, techniques that can be employed to counter this. While the authors have very successfully been able to render panpot stereo into ambisonic format, with good perceptual and acoustic efficiency advantages (and there are a variety of other techniques for this, e.g.[15]), the reverse process can be employed in an effort to capture and preserve ambience labelling information, We have found that the simplest way of engendering believability is to base a synthesised soundfield on an ambisonic wild track recording, and, crucially, much of the coherence of the ambience labelling information *of the background* seems to be preserved when decoded to asymmetric arrays such as 5.1. The need to specify background/context is perhaps one of the few important generalisations that can be made in respect of cartoonifying ambience labelling information, which as we have suggested is a highly task-specific process, but ambisonic techniques work well when background forms a significant part of the audio material, and convolution techniques can also be very successful. Our experience in this regard has shown it is thereafter relatively easy to introduce sound objects believably into the field with the minimum of processing for coherence, which is particularly relevant as the coherence of the background/context

appears to be destroyed if more than one ambisonic source is used. The process is easiest for relatively distant objects, but proximate objects require processing for the temporal and angular disparity between direct and early reflected sound. Conversely, we have found that synthetic (encoded) ambisonic material whose ambience labelling characteristics are not well defined often appears diffuse and difficult to localise but still conveys a definite sense of 'over-thereness'. This type of processing appears ideally suited to situations where audio presentation is used in support of visuals and the depicted audio material is beyond the visual field, where accurate and solid localisation might be distracting.

With regard to localisation generally, we know that 'what' processing can successfully override 'where' processing, and that the depiction of an audio environment need not concentrate particularly or achieve great accuracy in the presentation of whereness, provided that sufficient depth of illusion is afforded the whatness. However, we also know that the parameters of the perceptual significance of an object require particular attention, which may include features such as proximity. Clearly, while the traditional means of representing this parameter by manipulating the dry/reverberant mix is a very successful example of cartoonification for stereo displays, it is over-simplistic for surround/3D (and especially periphonic) systems. Here, the homogeneity of direct and reflected sound and the angular and temporal disparity between audio components provide the most important cues, and these must be manipulated coherently. The importance of early reflections in providing such cues are of course well known, and it is also well known (although rarely in these terms) that the perceptually significant elements of a sounding object's interaction with its immediate environment can be cartoonified through the presentation of the, approximately, six or seven earliest reflections. This accords well with the authors own experiments, but what is missing in most current implementations is the appropriate localisation (or at least spatial decorrelation for simply cartoonified spaces) of these early reflections. Proximate objects, for instance, are not characterised simply by predominantly dry sound, which would sound deeply unnatural and in any case tend to be localised in the sounding speaker, but by angularly and temporally disparate early reflections. There is a problem, however, because current systems deal very poorly with the proximate zone. The obvious reason for this is that it is notoriously difficult to solidly localise sound objects in front of the speaker array. Generally, depth-of-field is available only beyond the speakers, but this space provides opportunities for achieving simulated depth of significance in audio material.

Another key feature of real environments is that of occlusion. One generalisation is that distance objects are more likely to suffer occlusion (a strongly frequency-dependent process) by more proximate bodies, especially if movement is involved. Movement itself, often presented disappointingly simply by the manipulation of the amplitude distribution of a point source across the speaker array, seems to require similar attention to proximity, in that the coherent alteration of the spatial and temporal character of early reflections gives strong auditory cues. However, a very significant cartoonification process is available in this respect, because the character of movement is often apprehensible quite independently of the change in localisation, especially for relatively distant objects. One example might include the sound of movement itself, this being often the perceptually most significant cue. Furthermore, using a variety of time-varying rolling comb-filters applied to the upper portion of the spectrum of an object's sound output, it is possible to simulate audio patterns that are perceptually interpreted as movement. When this is used in conjunction with treatments for facingness (as this usually correlates with the direction of travel and therefore provides useful information), the authors have found that accurate simulation of object trajectory by closely defining interaural difference information becomes unnecessary.

We have previously introduced the notion that ambience labelling information is very strongly conveyed in the high-frequency (HF) portion of the soundfield. This provides a further cartoonification opportunity, because information relating to proximity, location, movement, behaviour and so on can be very successfully presented using HF cues alone. This permits efficiencies in signal processing, and furthermore allows significant efficiency in system design to be achieved. As an example of this, the authors have experimented with hybrid system where low frequencies are rendered with very low spatial 'accuracy', allowing the capabilities of the low-frequency (LF) portion of the sound reproduction system to be very efficiently employed, especially if the material presented contains a significant LF portion, and a hierarchy of spatial accuracy built up whereby the detail of ambience labelling information is fairly accurately and very successfully represented using relatively much larger numbers of cheaper HF drivers.

There are of course many other cartoonification regimes that allow the efficient representation of spatial parameters, but, as we have stressed, few generalised signal processing techniques. However, we do not of course present cartoonification as a signal-processing tool. Instead, it should be regarded as a philosophical approach. It is intended to define the starting point for exactly the kind of empirical approach that has lead to the rich variety of signal processing techniques available today. By defining the *texture* of the audio material, as we have defined it, that is by specifying the perceptually important parameters of the space in which audio objects do or purport to exist and the significant emergent properties of the interaction of sounding objects with physical aspects of that space, we believe that synthetic soundfields are easily achievable that are genuinely believable.

3. CONCLUSIONS

There is widespread agreement amongst composers that soundfields produced exclusively by electroacoustic means lack a 'sense of space', a sense of 'out-thereness' or 'envelopment'. We have proposed 3-D audio as a 'space' that is not the classical physical space but an information environment that we term perceptual space. In our auditory perceptual space we have a unique class of information about the 'what' and 'where' that we call ambience labelling information. Application of the concept of perceptual significance allows a process of cartoonification that may 'accurately' and efficiently portray a sound environment that incorporates the required elements of believability.

What we have proposed is that, for a 3D audio display to convey information efficiently, it is as necessary to pay technical attention to the rendering of background 'non-things' as it is to pay attention to the depiction of informational 'things'. In this way, a given display can convey a great deal more information without 'filling up' the perceptual foreground, than would otherwise be possible. This information-based compression is analogous to that employed by normal perceptual mechanisms.

We believe that the principles outlined in this paper provide the foundations for a comprehensive and systematic approach to the production of 'virtual reality', or, more properly, 'artificial reality' that effectively utilises the capacity for 'selective attention' which is clearly a crucial characteristic of human perception.

REFERENCES

[1] Lennox, P.P., Myatt, A. & Vaughan, J.M. (1999) From Surround to True-3D. *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction.* Rovaniemi, Finland, 10-12 April 1999, pp.126-135.

[2] Gibson, J.J. 1969. The Psychology of Representation. Unpubl. manuscript.

[3] Gibson, J.J. 1979. *The Ecological Approach to Visual Perception.* Houghton Mifflin, Boston.

[4] Bregman, A. 2001. Pers. comm.

[5] Millar, S. 1994. *Understanding and Representing Space: Theories and Evidence from Studies with Blind and Sighted Children.* Clarendon Press, Oxford.

[6] Dixon, N.F. 1981. *Preconscious Processing.* John Wiley & Sons, Chichester.

[7] Moray, N. 1969. *Listening and Attention.* Penguin Books, Harmondsworth.

[8] Atkinson, J. 1993. *A Neurobiological Approach to the Development of 'Where' and 'What' Systems for Spatial Representation in Human Infants.* In: Eilan, N., McCarthy, R. and Brewer, B. (eds.) *Spatial Representation. Problems in Philosophy and Psychology.* Blackwell, Oxford, pp.325-339.

[9] Gerzon, M.A. (1974) Surround-sound Psychoacoustics. Criteria for the design of matrix and discrete surround-sound systems. *Wireless World* 80, pp.483-486.

[10] Nielsen, S.H. 1993. Auditory Distance Perception in Different Rooms. *Journal of the Audio Engineering Society*, 41, 10, pp.755-770.

[11] Neuhoff, J.G. (2001) An Adaptive Bias in the Perception of Looming Auditory Motion. Unpubl. manuscript.

[12] Begault, D. (1999) Auditory and non-auditory factors that potentially influence virtual acoustic imagery. *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction.* Rovaniemi, Finland, 10-12 April 1999, pp.13-26

[13] Bamford, J.S. 1995. *An Analysis of Ambisonic Sound Systems of First and Second Order*. MSc Thesis, University of Waterloo, Ontario.

[14] Blumschein, E. 2001. Why and How to Revise the Traditional Theory of Auditory Function. Discussion Paper M30, Auditory Function Discussion Group (http://iesk.et.uni-magdeburg.de/~blumsche/, 12 January 2001).

[15] Gerzon, M.A. (1974) Surround-sound Psychoacoustics. Criteria for the design of matrix and discrete surround-sound systems. *Wireless World* 80, pp.483-486.

[16] Gerzon, M.A. 1977. *Some Production Facilities Available in Ambisonics.* NRDC Ambisonic Technology Report NRCD/FCC 4, November 1977.