VECIMS 2005 – IEEE International Conference on
Virtual Environments, Human-Computer Interfaces, and Measurement Systems
Giardini Naxos, Italy, 18-20 July 2005

Improving spatial perception through sound field simulation in VR

Regis Rossi A. Faria, Marcelo K. Zuffo, João Antônio Zuffo
Laboratory of Integrable Systems, Polytechnic School, University of Sao Paulo
Av. Prof. Luciano Gualberto, 158 tv3, 05508-900, Sao Paulo, SP, Brazil
Phone: +55-11-3091-9738, Fax: +55-11-3091-5665, Email: {regis, mkzuffo, jazuffo}@lsi.usp.br

Abstract *– A correct and wide coupling of sound to visual applications is still missing in most immersive VR environments, while future and advanced applications tend to demand a more realistic and integrated audiovisual solution to permit complete immersive experiences. Still there is a vast field of investigations till a correct and complete immersive system can reproduce realistic constructions of worlds. Sound fields simulation, although complex and of expensive implementation in the past, is now a potential candidate to improve spatial perception and correctness in CAVEs and other VR systems, but there are serious challenges and multiple techniques to do the job.*

In this paper we introduce our investigations in such fields and proposals to improve spatial perception and immersion experience in CAVEs through sound field simulation and correct matching of audio and visual cues. Additionally, a spatial sound immersion grading scale is proposed, to allow for system assessment and comparison of capabilities in delivering spatial immersion.

Keywords *– 3D sound, auralization, acoustic simulation, spatial perception*

I. INTRODUCTION

Most current immersive virtual reality systems and applications do not possess an efficient mechanism for correct spatial sound projection, capable of recreating a 3D sound field through multichannel auralization. In this terrain, more attention is routed for the visual sense. However, audiovisual applications more and more require that visual cues match aural cues, in order to increment the overall spatial perception. Popularization of multichannel systems presents new possibilities for sound field simulation, formats, techniques and speaker configurations.

Aiming at the integration of 3D spatial sound to immersive VR navigation has led us to several investigations towards the implementation of flexible and low-cost solutions for improving spatial sound impression in immersive audiovisual environments, such as CAVEs [1], liberating the creative power for new applications. Due to the nature of CAVEs, auralization seems to be a good candidate for sound field simulation and presentation on multichannel speaker layouts, dismissing headphones.

In this paper we present and discuss some of our investigations in this field. A brief overview of audiovisual perception is done, shed by our context. We then discuss spatial audio attributes that are important to quantify spatial perception in systems and applications and also to guide correct spatial sound system design. An immersion perception grading scale is proposed to measure the reached level of spatial sound immersion attained with existing systems. In the next sections current investigations towards sound field simulator design and implementations under way are then presented, and future directions are pointed out.

II. AUDIOVISUAL PERCEPTION

It is well known that visual perception is incredibly augmented by sound perception (and vice-versa), that correct assessment of distance and size is greatly improved by the presence of both mechanisms, and that they are complementary. Nature has throughout evolution provided ways in that one sense compensates for the lack of other, e.g. where one can not see backwards to notice danger approximations, but can listen and perceive sounds coming from the back.

Complete immersion perception in current VR systems depends not only on providing visual and aural outputs surrounding the users, but also on meeting a number of psychophysical requirements, such as correct correspondence of metrology characteristics of objects (e.g. shape, sizes, distances) on both visual and aural domains. This is not a trivial task, and is intimate connected to the scope and goals of pursuing realism in virtual reality (VR) applications.

Also, since vision and audition are in great part (if not most) neurological processes inside the brain, one cannot neglect that modeling inputs may have influence on this high level of perception process. In this paper, however, we are concerned only with the physical realization of virtual auditory worlds, addressing the acoustical component of such experience.

Cinema has through the years granted us with many trials and proofs of this, from the first break-point when sound was added, to the time when multichannel (surround) was introduced, presenting new challenges for our perceptual system to understand.

Psychoacoustics and artistic criteria were both used to set up a “standard” to display aural information in cinemas, such as allocation of voice and dialogs in front channels and special effects and movements to the side and surround channels. These made possible an undeniable improvement in spatial perception, and have undergone a kind of standardization, to be adopted by sound engineers in mastering movie sound tracks. However, this “standard” may be more a consequence of a commercial setup, constantly defining and shaping a media consumer culture, than specifically a standard for correct reproduction of the

audiovisual experience. This last has not been the real issue since the 1950's, not only because technology could not offer affordable multichannel infrastructure to make it possible in the past (as can today), but also because real reproduction of recorded audiovisual scenes was less desired than the ones artificially created. Illusions and surrealistic signs can be accomplished with simplifications in the models and technological tools. Art and science in this sense have always been influencing each other's evolution.

The current 5.1/6.1/7.1 surround standard [2] plays a special role in these conquests, and, due to its popularization in the last years, has gained attention from the scientific community, interested in making use of this setup to project finer and more correct sound fields, porting known auralization techniques and test new ones, making it the bridge towards new generations of "surround" technologies, named immersive.

We believe this is a trend for the future of audiovisual gears, and for this a line of investigations was proposed under the AUDIENCE project [3].

A. Spatial audio perception

The perception of spatiality in the aural domain is quite a simple experience to sense but a rather complex one to discriminate, quantify and classify. Sound quality is known to be a multidimensional phenomenon, and its complex structure has been addressed by several recent works [4,5,6].

Many previous works in this investigation arena had pointed out important attributes of sounds, of sounds sources, and of the environment, which relate directly to the perceived quality of spatiality or immersion in such environments. These naturally play an important role in establishing a mapping through which incremental levels of spatial perception can be quantified, and different situations can be compared.

Berg [4] has studied audio quality perception and proposed a method for systematic evaluation of perceived spatial quality. Zacharov [6] has addressed subjective mapping and characterization procedures for assessing spatial quality.

In these works the authors develop a comprehensive set of attributes and unravel the most relevant components related to the perception of spatial quality, opening ways for further proposals of techniques to measure spatial quality.

Several tools exist to create or explore spatiality in audio, both hardware and software solutions, and many more are constantly appearing in the market. Cost and application needs are the most effective constraint and requirement to define consumer and professional audio product quality. 3D is a trend, and different ends require more of a sense of direction and envelopment than a precise impression of real location of sound sources. Other applications may justify a more refined approach, where precise sound field perception is a must. We believe this is the case in complete immersive VR.

However, the level of "perfection" depends on the final application needs, which may in many cases use a simpler 3D sound technique or require a more robust and computer-expense technique. One needs a way to quantify how much impression of spatiality an application needs.

Berg and Zacharov identified a set of sound attributes to be important in spatial quality assessment, which we combined and present condensed in table 1.

Table I. Sound attributes

source width
ensemble width
source distance (distance to events)
localization (sense of direction)
source envelopment
room envelopment
room size
room level
room width
depth
sense of movement
frequency spectrum (low/high frequency content)
naturalness
presence (sense of space)
preference

These attributes basically emerge from the application of an evaluation method where the elicitation and structuring of personal constructs (descriptors proposed by subjects) are refined and clustered, until a stable set is achieved. The reader shall refer to [4] for a complete description of all attributes.

B. Sound immersion level scale

From Berg's and others' works and results, and based on the necessity for a simple mechanism to quantify the level of spatial quality, we propose a sound immersion level scale. The basic idea is to offer means of mapping attribute ratings to metrics of immersion capability of spatial audio systems.

Table II presents a 6 discrete (integer) sound immersion grading scale. This may be however conveniently adapted to a continuous scale. In this table, techniques for spatial sound generation are related to immersion levels and spatial perceptions. Besides the attributes in table I above, we consider also other characteristics to influence in the grading task, such as the audio quality (temporal definition, S/N, THD, timbre, and other figures of audio quality) and image quality (definition, localization).

The ITU-R BS.1116-1 standard [7], although not comprehensive in all the aspects covered in this paper, is a

reference guideline for subjective test procedures setup and execution. Although indispensable for practical assessments, these topics are beyond our purpose here.

Table II. Sound immersion level scale

level	techniques/methods	perceptions (results)
0	monaural "dry" signal	no immersion
1	reverberation, echoes	spaciousness, ambience
2	panning (between speakers), stereo, 5.1… (n.m surround multichannel)	direction, movement
3	amplitude panning, VBAP	correct positioning in limited regions
4	HRTF, periphony (Ambisonics, WFS, etc.)	stable 2D sound fields
5	HRTF, periphony (Ambisonics, WFS, etc.)	stable 3D sound fields, accurate distance and localization

Some premises are assumed prior using the above scale: a) sounds are reproduced artificially from discrete/point sources (speakers/transducers); b) speakers *mimic* or artificially reproduce analog original sound sources, through an indirect sign mediation, i.e., they "speak on behalf" of utility sound programs; and 3) one level incorporates previous level's features and capabilities (cumulative). Regardless of the technique employed in the acoustic modeling and reproduction of the sound, we are interested in quantifying the capability of a speaker array to deliver a perceivable (and measurable) amount of spatial quality, in terms of the attributes introduced in table I.

Immersion level 0 refers to a monaural "dry signal" irradiating from one speaker that (despite of having a physical direction and positioning within the auditory space) does not represent or reconstruct the real direction/position that the audio program (primary source) suggests.

A suggestion of spatiality (ambience) upgrades our sensations to level 1 of immersion, eliciting the experience of echoes and reverberation that take place in the "remote" world. Through these, the user can refer to the size and type of environment he is "aurally" invited into.

The next level of immersion (level 2) inherits previous level capabilities and additionally permits the perception of movements and the first cues for assessing direction in the reconstructed auditory scene. For the first time a larger area of the auditory place is used to map and (re)scale the virtual world and project it locally. Neher draws in [5] a simple sketch of an auditory environment identifying sound scene components and illustrating various spatial attributes graphically.

Level 3 permits a correct positioning, sense of distance and stable image formation for virtual sources in limited areas. Vector Based Amplitude Panning (VBAP) techniques, just to say, can deliver these results.

Level 4 permits the formation of a stable and more realistic 2D sound field. Pantophonic and periphonic techniques – such as Ambisonics [8, 9], Ambiophonics [10], and Wave Field Synthesis (WFS) [11] – are capable of delivering this level of spatial quality (and higher). Mapping of the virtual world onto local auditory area is more accurate. A minimum of 4 speakers is required, and phase synchronization between channels/speakers is more critical.

Failure to satisfy these requirements leads to unstable images, audible artifacts and distortion in the sound field. Quadraphonic systems from the 1980's aimed to reach this degree of spatial impression, but failed due to technical issues, both from misconception designs and hardware limitations. Dolby Surround and successors are, however, an exception, mainly because they defined specific perceptual goals to pursue in creating "surround effects" and improved the technology generation after generation, adapting it to the new digital medias, which are multichannel-capable in essence.

Level 5 will finally permit the synthesis of 3D stable images around the user, thus permitting his/her complete envelopment and taking him/her to any possible aural illusion, be it of a real (recorded) world or an artificial (virtual) one. Rendering of distance and localization are supposed to be as accurate as in ideally real world. This level is mostly associated with the employment of sophisticated acoustic modeling techniques and sound field simulators.

This scale not only provides means for a fast understanding of spatial perceived quality and how much to expect from an application or sound system, but also may meet market requirements for a standard way to inform the capabilities of their products and solutions. Also, it provides means for comparing spatial quality achieved in different system implementations, which is a frequent need when sound demonstrations are not at reach.

Although level 5 may be everyone's goal for marketing in the future, most applications for the consumer market (and in an affordable and satisfactory way) require levels 2 to 4, as in games. Level 5 may be of more importance for applications where precise simulation and reconstruction of real auditory scenes are strict requirements, such as in critical missions, VR training, engineering design, etc.

III. CURRENT INVESTIGATIONS

A. Sound field simulation in CAVEs

VRs applications essentially and naturally demand a more realistic realization of the auditory world than any other application, for obvious reasons. Auralization techniques seem to be good candidates to accomplish correct level-5 sound field simulations in immersive environments, and it is our current goal to implement and integrate them in the *CAVERNA Digital* – a 5-sided CAVE virtual reality system at

the University of São Paulo – for hosting advanced audiovisual applications that could not be possible before, without an improved sense of aural reality, visually matched. The CAVE concept was introduced by Cruz-Neira in 1992/93 [1], and its free-movement and multi-user nature suggests the usage of a multichannel audio approach to auralize it.

The AUDIENCE project is a research and development initiative seeking solutions for immersive audio in the CAVERNA Digital, aiming to the implementation of a flexible and scalable system for spatial (2D/3D) audio reproduction, attending applications that possess several sound formats, from stereo/bi-aural, commercial "surround" formats, up to advanced formats of 3D multi-channel audio coding and sound field simulation, as Ambisonics and WFS, with the flexibility of being able to modify the space configuration and the number of loudspeakers, depending on the auralization method. [3]

Higher sound immersion levels require more computational power to process complex sound scene descriptions, taking into account more complete scene attributes sets, and usually use more accurate acoustic models and rendering techniques, for low and high frequency ends. Some models and simulators are really impractical for real time applications provided that not enough computer power is available. A complete simulation of sound waves propagation by, for example, solving the wave equation involves high computational costs, and may be practical only with supercomputing resources, something not at reach of popular gears.

However, powered by a cluster computer system, we intend to investigate level-4 and level-5 simulations of sound fields coupled to visual navigation in the CAVERNA Digital, even considering the integration of complex models. This is expected to provide insights into another goal of the AUDIENCE project: the development of solutions for low cost auralizators, making use of commodity audio gears.

Currently we are investigating a multichannel auralization scheme using Ambisonics coding and decoding techniques [8]. Ambisonics is an elegant mathematical approach to register and reproduce a 3D sound field introduced by Gerzon in the 1970's, but did not reach popularization, mainly due to technology limitations. It requires (for a first order setup) only 4 channels – *x, y, z, w* – to complete encode a 3D sound field [9].

An Ambisonics decoder is then responsible to decode these signals and compute sound outputs for an array of speakers, which may vary in number and position. These last characteristics turn Ambisonics into a very flexible, scalable and interesting sound field simulator for several situations.

B. Building an auralizator

Audiovisual environments in VR are artificially constructed rather than recorded. Objects' aural attributes inside it are simulated to compute an acoustic realization of the sound propagation in the virtual world. The outputs of this simulation are then used to process "dry sounds" and generate a spatial (and temporal) representation for them (intermediate codification format). Spatial coded sounds are finally decoded and/or mapped to produce N loudspeaker outputs.

This is a general spatial audio production/rendering scheme, adequate for multichannel setups, and flexible enough to permit the use of different acoustic models, spatial sound codecs and players, as approached by Faria [12].

Figure 1 shows a block diagram for a sound field simulator following this scheme. The blocks at left refer to the sound synthesizer (sound sources) and the VR application, where the user interacts with a navigator and an acoustic scene model is described. The central block is responsible for the acoustic simulation and spatial sound codification, thus generating the spatial coded sound vectors. The block at right contain a mixer (when several sound sources are under simulation), a spatial sound decoder and the final mapper/player to speakers, which may also include additional filters for deconvolving speaker/room interferences and proper equalize the auditory space.

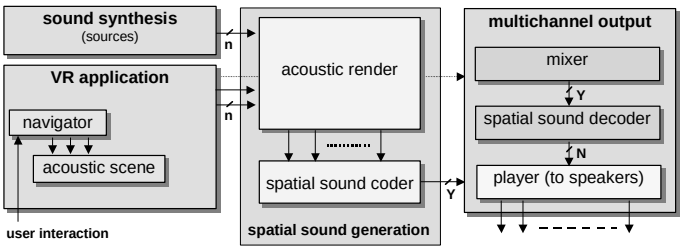

Fig. 1. Block diagram for a sound field auralizator

Both speakers' and virtual sources' coordinates respectively in the real and virtual worlds must be known before engaging sound field simulation.

Figure 2 below illustrates a virtual source and speakers having their location tracked in a CAVE sound field auralization setup.

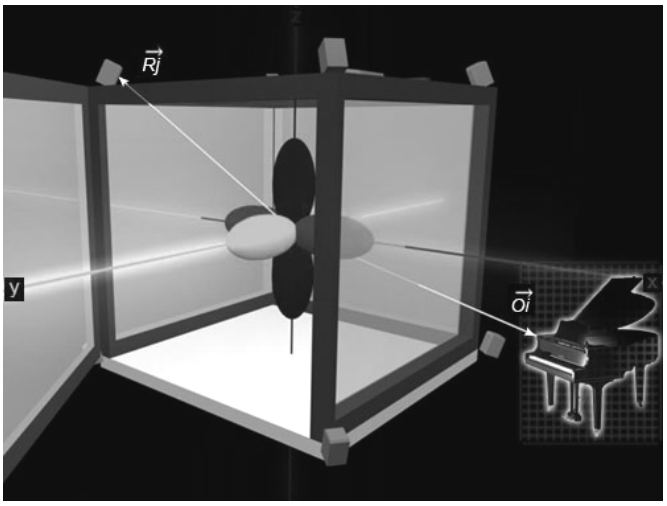

Fig. 2. Virtual sources and speakers in a CAVE sound field auralizator

We have designed an Ambisonics complete solution for CAVEs, and are mounting the first Ambisonics setup of up to 16 channels in the CAVERNA Digital.

Designing Ambisonics for CAVEs is a complex task. Audio processing is essentially a serial pipeline, collecting and propagating distortions and malformations throughout the channel. Main challenges include optimal speaker positioning, local acoustic compensation, and overall system calibration and synchronization, where aural and visual cues must match to provide correct audiovisual navigation. High quality speakers, amplifiers and cabling are also a must. These and details of implementation will be addressed in a future paper, as well as other sound field techniques, such as wave field synthesis, whose implementation in CAVEs will require the development of special drivers arrays.

For WFS, the forbidden area behind the screens (due to back optical projection) represents a challenge for its physical realization in CAVEs, since the ear's height is the best elevation to position speaker for correct azimuth perception. This, however, may force an architectural evolution in CAVEs and other immersive VR environments, requiring new transducer technology for sound, such as flat speaker panels.

C. Calibration and Experimental tests for improved spatial perception

The CAVERNA Digital is being sonorized by eight LANDO high-fidelity speakers, which can be mounted in several positions behind the screens and around the central auditory area. This will be upgraded to a 16-loudspeaker setup. Calibration tasks involve the proper deconvolution of the screen filtering and compensation for local acoustics interferences.

Experimental tests are planned to subjectively study spatial perception for several speaker configurations, from regular polygons (e.g. cube and octahedron) to irregular geometries, such as 5.1/7.1 surround positioning and others. Figure 3 shows one possible configuration, exploiting a cubic approach (surrounding the corners) plus reinforcement speakers.

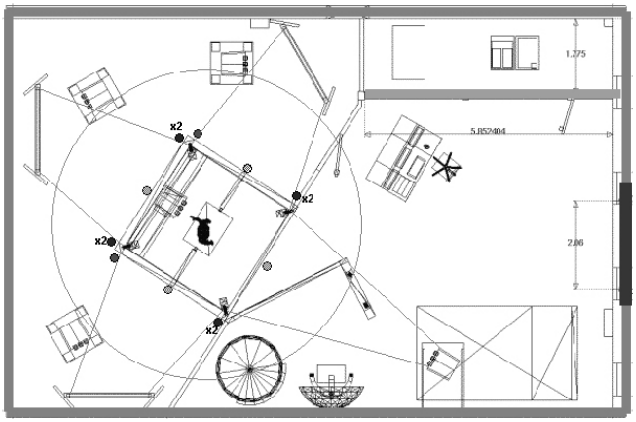

Fig. 3. A cubic (plus reinforcement) speaker configuration in the CAVE

IV. CONCLUSION AND FUTURE WORKS

Our methodology towards improved spatial perception in immersive VR includes first a multichannel setup to cover a set of 2D/3D audio solutions (software and hardware) and then the implementation and testing of 3D sound field generation and reproduction techniques, such as Ambisonics and WFS, coupled/integrated to visual applications, to pursue enhanced degrees of immersion experience.

The perception of spatial sound is addressed in several recent papers, most concerned with evaluation methodologies to measure spatial features consistently [4,5,6]. The idea developed here concerns a tool for spatial audio system classifying/grading in their capability of reproducing certain spatial attributes accurately and consequently their ability to project 2D and 3D sound spaces. A sound immersion scale was proposed as a reference tool to assess the spatial quality of 2D/3D sound systems, and to permit their categorization and/or comparison.

A series of calibration tasks are planned to obtain a correct set of parameters to control the sound synthesis and sound field production, so that aural cues fit in physical attributes to visual cues. This includes correct choice for acoustic attributes (for absorption, reflection, etc.) within the acoustic simulator, psychoacoustic weighting (frequency and amplitude) and pre- and post-processing parameters required to avoid saturation (clipping), to setup correct amplitude (sound pressure) for speakers, and to control local acoustic compensation.

The next tasks will encompass test applications, designed to permit a systematic assessment of the perceived spatial quality in the immersive audiovisual virtual environment.

A. Future works

An objective method or expression to calculate the immersion level of an audio system is desired and expected to be developed based on previous defined attribute scales and evaluation methods (proposed and discussed in several works). This is required to consistently map spatial attributes ratings to levels in the proposed sound immersion grading scale. It is important to notice that grading may also be modulated by the correct perception of the visual cues.

It is important to notice that general audio quality figures may lead the overall quality assessment up or down to some extension. For example, a 4.1 immersion level grading may fall behind 4.0 due to loss of spectral resolution or higher noise level, which could in theory disturb the perceived stability of a virtual sound image. Artifacts and lack of calibration might also contribute to a decrease in immersion level perception, and one shall carefully consider situations when minor faults have to be properly contained.

Additionally, objective acoustical metrics (such as reverberation time, energy decay, high/low frequency content, strength, and others) may contribute to establish a more formal, direct and less subjective mapping of spatial attributes to levels within the immersion grading scale.

The audio industry may benefit from such a methodology to quality assessment of products, both hardware and software, specially the game and home-theater industries.

Future measurements of perceptual cues in virtual worlds shall be addressed, through tests in virtual environments constructed with correct distance and situation perception for both acoustic and visual point of view to evaluate the fitness between visual and sound perception to the same object. This includes the evaluation of gestures, navigation, and influence of application usage in the perception of immersion to improve human interaction in projected virtual audiovisual worlds.

This is very significant if we want to propose a method or technique to calibrate sound and visual systems together, and make sound cues match visual cues.

ACKNOWLEDGMENT

The authors wish to thank the industrial partners of the AUDIENCE project, the “SISCOMPRO” project support by FINEP Brazilian agency, and all colleagues from the CAVERNA Digital at the University of Sao Paulo who have been providing technical support to the project.

REFERENCES

[1] C. CRUZ-NEIRA; D.J. SANDIN; T.A. DEFANTI. “Surround-screen projection-based virtual reality: The design and Implementation of the CAVE”. Proceedings of the SIGGRAPH 1993. ACM SIGGRAPH, Anaheim, July 1993

[2] International Telecommunications Union. “Recommendation ITU-R BS.775-1. Multichannel stereophonic sound system with and without accompanying picture”, 1994 (rev.1992).

[3] R. Faria. “AUDIENCE – Audio Immersion Experience by Computer Emulation Project”. http://www.lsi.usp.br/interativos/nem/audience/. 2004/2005.

[4] J. Berg and F. Rumsey. “Systematic evaluation of perceived spatial quality”. Proceedings of the AES 24th International Conference on Multichannel Audio, pp.184-198, Banff, 26-28 June 2003.

[5] T. Neher et al. “Unidimensional simulation of the spatial attribute ‘ensemble depth’ for training purposes, part 1: pilot study into early reflection pattern characteristics”. Proceedings of the AES 24th International Conference on Multichannel Audio, pp. 123-137, Banff, 26-28 June 2003.

[6] N. Zacharov and K. Koivuniemi. “Unravelling the perception of spatial sound reproduction: techniques and experimental design”. Proceedings of the AES 19th International Conference on Surround Sound, pp. 272-286, Schloss Elmau, 21-24 June 2001.

[7] International Telecommunications Union. “Recommendation ITU-R BS.1116-1. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems”, 1997.

[8] AMBISONICS.NET. http://www.ambisonic.net

[9] D.G. Malham and A. Myatt. “3-D sound spatialization using Ambisonics techniques”. Computer Music Journal , v.19, n.4, p.58-70, Winter 1995.

[10] Glasgal, R. “Ambiophonics”. http://www.ambiophonics.org/. 2004.

[11] D. De Vries and M.M. Boone. “Wave field synthesis and analysis using array technology”. Proceedings of the 1999 IEEE Workshop on applications of signal processing to audio and acoustics, pp.15-18. New Paltz, 17-20 October 1999.

[12] R.R.A. Faria. “Auralização em ambientes audiovisuais imersivos”. (Auralization in immersive audiovisual environments). PhD. Thesis in Electronic Engineering. Polytechnic School of the University of São Paulo. São Paulo, 2005.