Spatial Sound and Multimodal Interaction in Immersive Environments

Francesco Grani, Dan Overholt, Cumhur Erkut, Steven Gelineck, Georgios Triantafyllidis, Rolf Nordahl, and Stefania Serafin
Aalborg University Copenhagen
Department of Architecture, Design and Media Technology
A. C. Meyers Vænge 15, 2450 Copenhagen, Denmark
{fg, dano, cer, stg, gt, rn, sts} @create.aau.dk

ABSTRACT

Spatial sound and interactivity are key elements of investigation at the Sound And Music Computing master program at Aalborg University Copenhagen.

We present a collection of research directions and recent results from work in these areas, with the focus on our multifaceted approaches to two primary problem areas: 1) creation of interactive spatial audio experiences for immersive virtual and augmented reality scenarios, and 2) production and mixing of spatial audio for cinema, music, and other artistic contexts. Several ongoing research projects are described, wherein the latest developments are discussed.

These include elements in which we have provided sonic interaction in virtual environments, interactivity with volumetric sound sources using VBAP and Wave Field Synthesis (WFS), and binaural sound for virtual environments and spatial audio mixing. We show that the variety of approaches presented here are necessary in order to optimize interactivity with spatial audio for each particular type of task.

Categories and Subject Descriptors

H.5.1 Multimedia Information Systems [**H.5.2 User Interfaces**]: H.5.5 Sound and Music Computing.

General Terms

Design, Experimentation.

Keywords

Multimodal interaction, Virtual environments, Spatial sound, Binaural sound, Wave field synthesis.

1. INTRODUCTION

In the following paragraphs, a review of the state of the art on interaction with spatial sound delivered through the use of arrays of loudspeaker is presented.

One of the early works that considered the importance of a users point of view in relation to the acoustical perception of the interactive spatial sound field, is by Melchior et al. [26]. In this work, an augmented reality system is coupled with an array of loudspeakers to deliver a wave field synthesis generated sound field, in which users could interact perceiving his own visual perspective of the given scene. A framework is proposed for possible tasks to be performed with such a system, as: room simulation, sound filtering, spatial layout of sound sources, and sound editing, and a simple prototype in which two users can work simultaneously on the same auditory scene is implemented and discussed by authors.

Another system capable of hosting multiple users, is the system proposed by Springer et al. [38] that combines a projection-based multi-viewer stereo display and wave field synthesis to simulate spatial sound sources of various kinds. Thanks to the absence of a sweet spot given by the wave field synthesis sound rendering, multiple users at the same time can participate to the interactive experience, interacting with virtual sonic objects that could be manipulated in real time. The authors' choice of a 2D multi-viewer stereo display relies also on the work done by Melchior et al. [25, 24] where it is suggested that high quality on the acoustic perspective can be achieved through the use of wave field synthesis in the particular case of 2D projection, as the image observed by viewers will match consistently the spatial depth of sound given by WFS. Although no strict user evaluation has been performed in this work, results from observations showed that users perceived a good level of visual and auditory consistency of the virtual objects, as well as a high level of natural interaction.

In a recent work by [30] a mid-air direct interaction system is proposed that allows users to manipulate and "touch" sound sources in real time. A wave field synthesis system is employed to deliver spatial sound, in a combination with a marker-based motion tracking system that captures the position of users hands to superimpose focused sound sources and at the locations of hand gestures. Gestures are used to trigger sound-manipulating actions such as: move, pick-up, place, and release a sound source. This approach allows users to interact with sounds in a natural way, without the use of hand-held button controllers or visual interfaces. To acheive some perception of height localization (which their WFS system was not designed for), the authors relied on the use of the ventriloquist effect, obtained trough the use of physical objects such as a bottle, which could be "filled"

AM '15, October 7-9, 2015, Thessaloniki, Greece

DOI: 10.1145/2814895.2814919

with a sound spilled into it through a "drop gesture" performed by users. The sound in the bottle could then be "heart", moved around, and placed on different tables. Following the encouraging results obtained by users evaluation in localization accuracy, the authors physically built a spatial mixing room, in which sound objects can be manipulated and placed in precise spots around the listener.

A previous real-time gesture control interface for wave field synthesis was proposed in the work presented by Fohl et al. [9], in which users could trigger sound-manipulating actions, through the use of three elementary gestures: select/deselect pose, and circular/radial movement. As later done in the Boom Room experiment, this work capitalises on hand tracking, with optical tracking via a standard 3-D marker placed on one of the user's hands. Two of the three analysed gestured (the select/deselect pose and the circular movement) showed a high level of affordance, as users found a natural match between their performed gestures and the expected results. However, the radial motion gesture showed some problems, as it was often difficult for the users to control the radial motion of a sound source, especially when pulling the source nearer. The authors point out how the acoustical perception of a source position on the radial axis is quite difficult and propose better calibration and fine-tuning of gesture tracking as a possible solution to cope with this issue.

2. SONIC INTERACTION IN VIRTUAL ENVIRONMENTS

We investigate how novel interactions with sound can be used for virtual reality applications. To achieve this goal, we designed several interfaces embedded with sensors and actuators, that drive physics-based synthesis models. We are interested in using these interfaces to allow users to naturally interact in virtual environments as they interact in the real world. As an example, we have been working with walking interaction in virtual environments, building a pair of shoes embedded with sensors and actuators. The sounds are spatialized in the laboratory using the vector based amplitude panning technique (VBAP) [36]. The correct positioning of the virtual sound sources according to the actual user position is achieved by tracking the user shoes using a motion capture synthesis, and mapping the position of the user to the VBAP algorithm. We noticed that users perceive that the footstep sounds are emanating at feet level, instead of coming from the speakers. This is an analogy to the ventriloquism effect in the audio-tactile domain. More information can be found here: [10].

3. INTERACTION AND WAVEFIELD SYNTHESIS

3.1 Spatial Extent

Synthesis of volumetric virtual sources is a useful technique for sonic interaction in virtual environments. This task can be simplified to the synthesis of spatial extent and carried out perceptually or physically. Previous research in Directional Audio Coding has shown that spatial extent can be perceptually synthesized with monophonic sources by applying a time-frequency-space decomposition, then randomly or structurally distributing time-frequency bins of the source signal [37, 19, 34]. Similar principles apply to the physical techniques, most notably to the Wave Field Synthesis, without signal decompositions however. In WFS, the synthesis rather relies on the randomized or optimally phase-delayed combination of elementary virtual sound sources or spherical harmonics [1] [5].

These methods can equally perform well on recorded or synthesized source signals. If the task involves ecological synthetic sounds as a result of group activity, such as clapping, walking, and other sound-producing events, combining physically based sound and spatial extent synthesis into a single system might be profitable in terms of computational cost, perceived sound quality and interaction fidelity. In WFS, such an approach has been tried with idealized source models [1, 29], but not with ecological models to our knowledge.

4. BINAURAL SOUND

4.1 Binaural Sound in VE

Although research on spatialized sound rendering has a relative long history in the computer music and sound processing community [4, 18, 2], little research has investigated the effect of spatialized sound when included in a virtual reality environment. An exception is the work presented in [20], where two experiments were proposed in order to investigate potential benefits of high quality auditory rendering in virtual reality. Results showed that the condition with high quality auditory rendering elicits a higher sensation of presence. Also, previous research integrating a CAVE-like environment with high quality sound rendering has been focussing on sound delivered through surround sets of speakers [16].

Research has been conducted within our group [14, 15] to investigate users' appreciation of binaural sound rendering created using non-personalized head-related transfer functions (HTRF). In a wide four-sided CAVE environment where users were allowed to walk inside a natural virtual environment, we compared a binaurally spatialized sound scene in which the location of sonic objects spatially matched the location of the corresponding visual objects, with an inconsistent binaural sound scene where the location of sonic objects was spatially incongruous with the location of visual objects, and with a standard stereo sound scene. Results of walking experiment showed increased preference ratings for the consistent binaural audio rendering. As expected incongruent spatial cues were ranked significantly lower.

A subsequent experiment compared then how different "attractors" (audio and/or visual, static or dynamic) modify the user's attention while walking in a VE, to provide possible guidelines to the design of virtual attractors [32, 33]. The results of the conducted experiment showed how audio-visual attractors are the most efficient attractors in order to capture users' attention toward the inside of the CAVE.

4.2 Spatial Audio Mixing

Apart from utilising spatial sound for VEs we are exploring the growing interest in production of spatial audio for cinema, music and other artistic contexts [28, 23]. We have been investigating how physical user interfaces can help content creators mix for spatial audio, extending beyond the traditional channel strip based mixing console—both for stereo production [11] and for 3D spatial audio [12]. Instead of adjusting volume and panning of the sound source using faders and rotary knobs, the basic approach is to adjust distance and angle(s) in reference to a virtual listening position as if positioning sound sources in a virtual environment.

Exploratory studies suggest that while there is great potential in utilising this control metaphor for mixing, there are several challenges involved in implementing it into a professional mixing context. The most prominent challenges include clutter and lack of overview when the number of channels increase, gaining fast access to underlying audio effects parameters, the lack of tactile and haptic feedback (when using multitouch or mid-air hand gestures [12]), lack of precision, and too high dependancy on visual feedback. We approach these challenges from a user centred Human Computer Interaction (HCI) point of view, developing and testing alternative prototypes that implement solutions to said challenges—for instance by extending multitouch surfaces with smart tangibles for improved tangibility or extending GUIs with graphical layers for reduced clutter [13]. Currently, we are exploring interfaces for mixing binaural audio, which explore active haptic feedback for faster and more precise interaction. Building on related work[27] we test different alternative forms of haptic feedback for faster and more precise selection and manipulation of virtual sound sources in 3D exploring whether it is possible to reduce the dependancy on visual feedback through haptics.

5. MOTION SENSOR TECHNOLOGIES

Motion sensor technology is the discipline that processes, digitalizes, and detects the position and/or velocity of people and objects in order to interact with software systems. it has been establishing itself as one of the most relevant techniques for designing and implementing a Natural User Interface (NUI). NUIs are human-machine interfaces that enable the user to interact in a natural way with software systems. The goals of NUIs are to be natural and intuitive.

In this context, there are numerous devices that try to act as motion sensors. The first breakthrough was the Wii Remote in 2006, which was the primary controller for Nintendo's Wii console. A main feature of the Wii Remote is its motion sensing capability, which allows the user to interact with and manipulate items on screen via gesture recognition and pointing through the use of accelerometer and optical sensor technology. In a similar context, PlayStation Move was a motion-sensing game controller platform by Sony Computer Entertainment (SCE), first released for the PlayStation 3 (PS3) video game console in 2009. Based around a handheld motion controller wand, PlayStation Move uses inertial sensors in the wand to detect its motion, and the wand's position is tracked using a PlayStation webcam.

The next breakthrough was Kinect, which is a line of motion sensing input devices by Microsoft for Xbox 360 and Xbox One video game consoles and Windows PCs. Based around a webcam-style add-on peripheral, it enables users to control and interact with their console/computer without the need for a game controller, through a natural user interface using gestures (and spoken commands). Kinect for Xbox which was launched in November 2010 and its launch was indeed a success: it was and it is still a break-through in the gaming world and it holds the Guinness World Record for being the "fastest selling consumer electronics device" ahead of the iPhone and the iPad. In December 2010, PrimeSense released a set of open source drivers and APIs for Kinect that enabled software developers to develop Windows applications using the Kinect sensor. Finally, on June 17 2011 Microsoft launched the Kinect SDK beta, which is a set of libraries and APIs that enable us to design and develop software applications on Microsoft platforms using the Kinect sensor as a multimodal interface. With the launch of the Kinect for Windows device and the Kinect SDK, motion control computing is now a discipline that everyone can shape easily, writing simple and powerful software applications.

A new addition to this list of the widely used motion sensing devices is the Leap Motion. Leap Motion is sensor device that supports hand and finger motions as input, analogous to a mouse, but requiring no hand contact or touching. The device started full-scale shipping in July 2013. Leap Motion is using two monochromatic IR cameras and three infrared LEDs, the device observes a roughly hemispherical area, to a distance of about 1 meter. The LEDs generate a 3D pattern of dots of IR light and the cameras generate almost 300 frames per second of reflected data, which is then sent through a USB cable to the host computer, where it is analyzed by the Leap Motion controller software, in some way synthesizing 3D position data by comparing the 2D frames generated by the two cameras.

The final addition to such sensing devices is Myo. Myo lets you control a computer with gestures. But unlike the previous devices, which rely on optical sensors, Myo employs a combination of motion sensing and muscular activity. The actual MYO device is an armband. When worn, it senses gestures, and sends the corresponding signal (via Bluetooth 4.0) to a paired device.

5.1 Gesture recognition for sound interaction and creation

Gesture Recognition techniques using motion sensor technologies have been utilized in applications varying from controlling multimedia playback [22], to web browsing [21], to medical imaging interaction [17] and others [7]. On the aforementioned examples, the interaction with the user does not produce multimedia or sound content, instead it affects the way the content is being represented. As a result, there are not strict timing constraints and Gesture Recognition can be used to provide inputs in similar scenarios. In the field of real-time sound content creation, Odowichuk et. al. [31] have created a platform that simulates a specific music instrument (xylophone) playing. This is achieved by, initially creating a virtual representation of the instrument, then extracting motion characteristics from the gestures, spatially mapping them and parse the resulting values as parameters to a sound generator.

Churnside et. al. [6] designed a gestured-based audio interface system that uses Joint Coordinates to adjust the speed and volume of audio and video playback.

Several low-latency techniques are used for human-robot interaction, but they address to trained operators and assume specific underlying hardware setup [39], [3], [41].

On more generic frameworks, Deshayes et.al. have developed a framework for gesture-based applications, with Statechart modeling [8]. Even though their approach provides solutions on the application development and verification side, they are using traditional Gesture Recognition techniques.

In this context, our research focuses on examining the sensors to be used as a multimodal interface, such as Voice and

Gestures [40] and for real-time multimedia content production [35].

6. CONCLUSIONS

In this paper we presented an overview of the interactive gestural driven spatial sound technologies used at XXX. We are experimenting with these interfaces from several perspectives, from the point of view of use in sonic interaction design and interactive multimodal interfaces, to their potential in creating immersive virtual environments where a sense of presence in achieved. As the paper shows, we are interested in spatial sound from different perspective: from an engineering perspective to improve the state of the art of physics based simulations, to a computer scientist perspective to implement efficient real-time algorithms, to an interaction designer perspective to understand how to create gestural based interfaces for these algorithms, and also an human centered perspective to understand sense of immersion and presence when spatial sound is combined with multi sensory environments, for example for virtual reality applications.

7. REFERENCES

[1] Jens Ahrens and Sascha Spors. Two physical models for spatially extended virtual sound sources. In *Proc. AES Convention*, New York, NY, USA, 2011.

[2] V Ralph Algazi, Richard O Duda, Dennis M Thompson, and Carlos Avendano. The cipic hrtf database. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 99–102. IEEE, 2001.

[3] P. Barattini, C. Morand, and N.M. Robertson. A proposed gesture set for the control of industrial collaborative robots. In *RO-MAN, 2012 IEEE*, pages 132–137, Sept 2012.

[4] Durand R Begault et al. *3-D sound for virtual reality and multimedia*, volume 955. Citeseer, 1994.

[5] Jung-Woo Choi. Extension of perceived source width using sound field reproduction systems. In *ICA 2013 Montreal*, 2013.

[6] Anthony Churnside, Chris Pike, and Max Leonard. Musical movements—gesture based audio interfaces. In *Audio Engineering Society Convention 131*, Oct 2011.

[7] L. Cruz, D. Lucio, and L. Velho. Kinect and rgbd images: Challenges and applications. In *Graphics, Patterns and Images Tutorials (SIBGRAPI-T), 2012 25th SIBGRAPI Conference on*, pages 36–49, Aug 2012.

[8] Romuald Deshayes and Tom Mens. Statechart modelling of interactive gesture-based applications. In *Proc. First International Workshop on Combining Design and Engineering of Interactive Systems through Models and Tools (ComDeis-Moto),. Lisbon, Portugal (September 2011), iNTERACT*, 2011.

[9] Wolfgang Fohl and Malte Nogalski. A gesture control interface for a wave field synthesis system. In *International Conference on New Interfaces for Musical Expression*, 2013.

[10] Federico Fontana and Yon Visell. *Walking with the Senses: Perceptual Techniques for Walking in Simulated Environments*. Logos-Verlag, 2012.

[11] Steven Gelineck, Morten Büchert, and Jesper Andersen. Towards a more flexible and creative music mixing interface. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 733–738. ACM, 2013.

[12] Steven Gelineck and Dannie Korsgaard. An exploratory evaluation of user interfaces for 3d audio mixing. In *Audio Engineering Society Convention 138*. Audio Engineering Society, 2015.

[13] Steven Gelineck, Dan Overholt, Morten Büchert, and Jesper Andersen. Towards an interface for music mixing based on smart tangibles and multitouch. In *Proc. of NIME*, 2013.

[14] Francesco Grani, Ferran Argelaguet, Valérie Gouranton, Marwan Badawi, Ronan Gaugne, Stefania Serafin, and Anatole Lecuyer. Design and evaluation of binaural auditory rendering for caves. In *Virtual Reality (VR), 2014 iEEE*, pages 73–74. IEEE, 2014.

[15] Francesco Grani, S Serafin, F Argelaguet, V Gouranton, M Badawi, R Gaugne, and Anatole Lécuyer. Audio-visual attractors for capturing attention to the screens when walking in cave systems. In *VR Workshop: Sonic Interaction in Virtual Environments (SIVE), 2014 IEEE*, pages 3–6. IEEE, 2014.

[16] Matti Gröhn, Tapio Lokki, and Tapio Takala. Localizing sound sources in a cave-like virtual environment with loudspeaker array reproduction. *Presence: Teleoperators and Virtual Environments*, 16(2):157–171, 2007.

[17] Rose Johnson, Kenton O'Hara, Abigail Sellen, Claire Cousins, and Antonio Criminisi. Exploring the potential for touchless interaction in image-guided interventional radiology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3323–3332, 2011.

[18] Jean-Marc Jot. Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia systems*, 7(1):55–69, 1999.

[19] Mikko-Ville Laitinen, Tapani Pihlajamäki, Cumhur Erkut, and Ville Pulkki. Parametric time-frequency representation of spatial sound in virtual worlds. *ACM Transactions on Applied Perception*, 2012.

[20] Pontus Larsson, Daniel Vastfjall, and Mendel Kleiner. Better presence and performance in virtual environments by improved binaural sound rendering. In *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*. Audio Engineering Society, 2002.

[21] Daniel Liebling and Meredith Ringel Morris. Kinected browser: Depth camera interaction for the web. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, pages 105–108, 2012.

[22] Shih-Yao Lin, Yun-Chien Lai, Li-Wei Chan, and Yi-Ping Hung. Real-time 3d model-based gesture tracking for multimedia control. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3822–3825, Aug 2010.

[23] Justin Mathew, Stéphane Huot, and Alan Blum. A morphological analysis of audio objects and their control methods for 3d audio. In *Proc. of NIME*, 2014.

[24] Frank Melchior, S Brix, and D De Vries. Zur kombination von wellenfeldsynthese mit monoskopischer und stereoskopischer bildwiedergabe. *FORTSCHRITTE DER AKUSTIK*, 31(1):207, 2005.

[25] Frank Melchior, Sandra Brix, Thomas Sporer, Thomas Roder, and Beate Klehs. Wave field syntheses in combination with 2d video projection. In *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. Audio Engineering Society, 2003.

[26] Frank Melchior, Tobias Laubach, and Diemer De Vries. Authoring and user interaction for the production of wave field synthesis content in an augmented reality system. In *Proceedings of the 4th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 48–51. IEEE Computer Society, 2005.

[27] Frank Melchior, Chris Pike, Matthew Brooks, and Stuart Grace. On the use of a haptic feedback device for sound source control in spatial audio systems. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[28] Frank Melchior and Sascha Spors. Spatial audio reproduction: from theory to production. In *tutorial, 129th Convention of the AES*, 2010.

[29] Alexander Müller and Rudolf Rabenstein. Physical Modeling for Spatial Sound Synthesis. In *Proc. Intl. Conf. Digital Audio Effects (DAFx)*, 2009.

[30] Jörg Müller, Matthias Geier, Christina Dicke, and Sascha Spors. The boomroom: mid-air direct interaction with virtual sound sources. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 247–256. ACM, 2014.

[31] G. Odowichuk, S. Trail, P. Driessen, W. Nie, and W. Page. Sensor fusion: Towards a fully expressive 3d music control interface. In *Communications, Computers and Signal Processing (PacRim), 2011 IEEE Pacific Rim Conference on*, pages 836–841, Aug 2011.

[32] Tabitha C Peck, Henry Fuchs, and Mary C Whitton. Evaluation of reorientation techniques and distractors for walking in large virtual environments. *Visualization and Computer Graphics, IEEE Transactions on*, 15(3):383–394, 2009.

[33] Tabitha C Peck, Henry Fuchs, and Mary C Whitton. Improved redirection with distractors: A large-scale-real-walking locomotion interface and its effect on navigation in virtual environments. In *Virtual Reality Conference (VR), 2010 IEEE*, pages 35–38. IEEE, 2010.

[34] Tapani Pihlajamäki, Olli Santala, and Ville Pulkki. Synthesis of Spatially Extended Virtual Source with Time-Frequency Decomposition of Mono Signals. *Journal of the Audio Engineering Society*, 62(7/8):467–484, July/August 2014.

[35] E. Potetsianakis, E. Ksylakis, and G. Triantafyllidis. A kinect-based framework for better user experience in real-time audiovisual content manipulation. In *Telecommunications and Multimedia (TEMU), 2014 International Conference on*, pages 238–242, July 2014.

[36] Ville Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

[37] Ville Pulkki, Mikko-Ville Laitinen, and Cumhur Erkut. Efficient spatial sound synthesis for virtual worlds. In *Proc. AES Intl. Conf*, London, UK, 2009.

[38] Jan P Springer, Christoph Sladeczek, Martin Scheffler, Jan Hochstrate, Frank Melchior, and Bernd Fröhlich. Combining wave field synthesis and multi-viewer stereo displays. In *Virtual Reality Conference, 2006*, pages 237–240. IEEE, 2006.

[39] M. Van den Bergh, D. Carton, R. de Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, and M. Buss. Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In *RO-MAN, 2011 IEEE*, pages 357–362, July 2011.

[40] N. Vidakis, M. Syntychakis, G. Triantafyllidis, and D. Akoumianakis. Multimodal natural user interaction for multiple applications: The gesture - voice example. In *Telecommunications and Multimedia (TEMU), 2012 International Conference on*, pages 208–213, July 2012.

[41] Dan Xu, Yen-Lun Chen, Chuan Lin, Xin Kong, and Xinyu Wu. Real-time dynamic gesture recognition system based on depth perception for robot navigation. In *Robotics and Biomimetics (ROBIO), 2012 IEEE International Conference on*, pages 689–694, Dec 2012.