51. IWK

Internationales Wissenschaftliches Kolloquium
International Scientific Colloquium

PROCEEDINGS 11-15 September 2006

FACULTY OF ELECTRICAL ENGINEERING AND INFORMATION SCIENCE

INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING - DEVICES AND SYSTEMS, MATERIALS AND TECHNOLOGIES FOR THE FUTURE

Startseite / Index:
http://www.db-thueringen.de/servlets/DocumentServlet?id=12391

TECHNISCHE UNIVERSITÄT
ILMENAU

Impressum

Herausgeber: Der Rektor der Technischen Universität Ilmenau
Univ.-Prof. Dr. rer. nat. habil. Peter Scharff

Redaktion: Referat Marketing und Studentische Angelegenheiten
Andrea Schneider

Fakultät für Elektrotechnik und Informationstechnik
Susanne Jakob
Dipl.-Ing. Helge Drumm

Redaktionsschluss: 07. Juli 2006

Technische Realisierung (CD-Rom-Ausgabe):
Institut für Medientechnik an der TU Ilmenau
Dipl.-Ing. Christian Weigel
Dipl.-Ing. Marco Albrecht
Dipl.-Ing. Helge Drumm

Technische Realisierung (Online-Ausgabe):
Universitätsbibliothek Ilmenau
ilmedia
Postfach 10 05 65
98684 Ilmenau

Verlag:
Verlag ISLE, Betriebsstätte des ISLE e.V.
Werner-von-Siemens-Str. 16
98693 Ilrnenau

ISBN (Druckausgabe): 3-938843-15-2
ISBN (CD-Rom-Ausgabe): 3-938843-16-0

Startseite / Index:
http://www.db-thueringen.de/servlets/DocumentServlet?id=12391

51st Internationales Wissenschaftliches Kolloquium
Technische Universität Ilmenau
September 11 – 15, 2006

Ulrich Reiter, Institut für Medientechnik, TU Ilmenau

TANGA Updated – A Modular Framework for Real Time Audio Rendering of Object-Based (MPEG-4) Audio Visual Scenes

INTRODUCTION

At the Institute of Media Technology at Technische Universität Ilmenau, we have designed an interactive MPEG-4 player (I3D) capable of rendering audio visual scenes during the last years [1]. The system is capable of rendering three-dimensional virtual scenes with embedded 2D video streams of arbitrary shape. It allows for interaction with the user and it provides a modular audio engine for real time rendering of (room) acoustic simulations and effects. This range of capabilities makes it unique among the MPEG-4 players available today.

Room acoustic simulations are very cost-intensive in terms of computing power, which makes interactive real time applications like these especially demanding. Due to the limited amount of computing power available in these systems, the simulation process has to be simplified. It is especially interesting to investigate the amount of simplification that goes unnoticed regarding the perceived overall quality of such audio visual scenes [2]. The I3D is used as a tool for these investigations [3]. In this paper, the current status of the audio engine TANGA (The Advanced Next Generation Audio) and a number of features of the I3D player are being discussed in detail.

MPEG-4 AUDIO

In MPEG-4 Audio, a number of different approaches to sound and room acoustic rendering are defined. Among other differences, they vary in complexity. The simplest approach is the so-called *spatialization* of sound sources: sound sources are panned in the virtual scene so that their visual and auditive representations coincide. Sound is coming from a defined location in the scene. Inside of rooms, it is most desirable not to render dry audio signals, but to provide the user with some kind of room acoustic simulation. The MPEG-4 Audio standard offers two methods for room acoustic rendering: The *Physical Approach* is based on the geometry of the virtual room, on the

materials used for the walls and the objects in that room, and on the location of sound source and listener. The second method, called *Perceptual Approach*, is based on a set of measurable acoustic criteria which are related to psychoacoustical parameters describing the room's acoustics [4]. These parameters ultimately serve to shape the impulse response of the virtual room, independently from the geometrical and material data.

TANGA RENDERING ENGINE

In the TANGA engine, all these methods have been implemented. Because the rendering method to be used for each sound source in a virtual room is defined in the MPEG-4 scene description, the I3D player interprets this scene description upon loading of the scene. The linking element between the parser of the scene description and the TANGA engine is the so-called TANGA mediator, which is responsible for scanning the scene description for keywords related to audio rendering. The TANGA mediator is then constructing a *Component Graph* from the information contained in the scene description which solely relates to the audio rendering process. Fig. 1 shows a simple example Component Graph for two *spatialized* sound sources (no room acoustic rendering, but correct panning of audio sources).

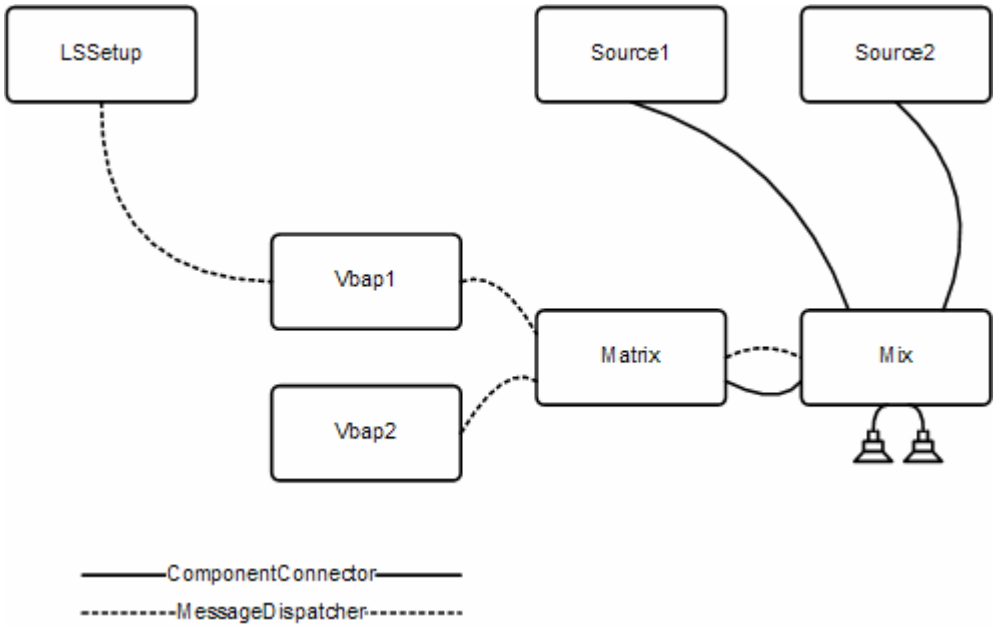

Fig. 1: Example of a very simple Component Graph. It provides correct panning of two sound sources for an arbitrary loudspeaker setup (defined in TangaLSSetup). Audio is output to the loudspeakers from the TangaMix Component.

Components are connected via so-called *Component Connectors* for the transmission of

audio and control data or via *Message Dispatchers* which are used to transmit only control data but no audio.

The system uses the 'PortAudio' API for audio input and output to any multichannel sound card and supports many different drivers on different platforms, e.g. ASIO and DirectSound onWindows and ALSA on Linux [11]. Latency of the TANGA System depends on the TANGA Components themselves, but is dominated by the time necessary to calculate the early reflections and late reverberation parts of a scene's room impulse response (RIR).

The *TangaEngine* class is defined as an abstract interface and hides the details of the underlying audio API to the rest of the system. This interface is currently implemented through the *PaTangaEngine* class, which provides a 'PortAudio' based render engine. It could be replaced with any audio API that provides some means of real time audio output.

One of the most important requirements for the TANGA Engine is that it should provide a DAC output timestamp. This should be the time when the samples being buffered will be played at the audio output, which is essential for synchronization purposes. 'PortAudio' was chosen because it provides such a timestamp and has in general very good real time support. Whereas 'PortAudio' also provides audio streams in blocking read/write mode, this feature is not useful for the TANGA Engine and we rely on the non-blocking audio streams which use a callback function for filling the output buffers.

This callback function invoked by 'PortAudio' is used to control the Tanga Engine, since 'PortAudio' ensures that this function is always called in time such that the output buffers are filled as needed by the audio hardware.

IMPLEMENTATION OF PHYSICAL APPROACH

For the MPEG-4 Audio *Physical Approach*, which includes room acoustic simulation, a mirror source method has been chosen to calculate the early reflections, see fig. 2. The late reverberation part is rendered using nested all pass filters. The complexity of the system is fully scalable, so the number of output channels as well as the maximum order of mirror sources to be calculated can be varied, among other parameters.

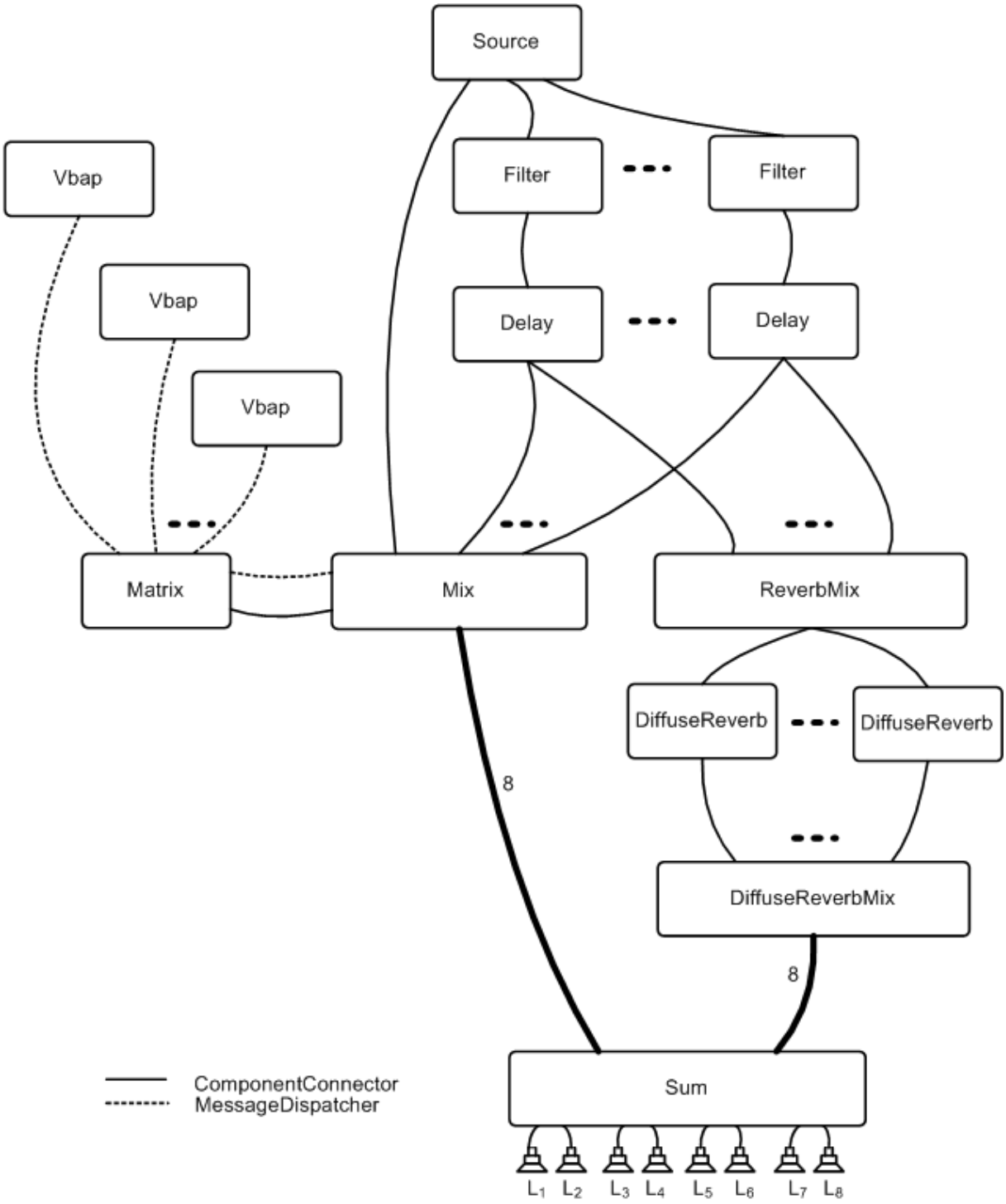

Fig. 2: A Component Graph representing the MPEG-4 Audio Physical Approach for a single sound source.

Early reflections are generated using *Filter/Delay* Components connected in parallel. The total number of early reflections to be simulated can be determined upon start of the I3D and solely depends on the performance of the rendering host / the PC. Therefore, on fast machines a higher number of early reflections can be rendered without changing the underlying MPEG-4 scene description.

The early reflections are also used as input signals for the generation of the diffuse

reverberation tail as described by Gardner [10]. They are summed in the *ReverbMix* Component and then a number of uncorrelated reverberation signals are created in the *DiffuseReverb* Components. These Components basically represent a Feedback Delay Network (FDN) each. For the rendering of the diffuse reverberation, the amount of computation power necessary for the calculus is proportional to the number of loudspeakers used for the reproduction of audio. Each loudspeaker needs to be fed with its own uncorrelated reverberation signal. Fig. 2 shows an example overview over the flexible implementation of the *Physical Approach*.

Special care needs to be taken to adjust the levels of the early reflections and the diffuse reverberation in such a way that they form one well sounding reverberator. As these are generated separately, there is a risk of gaps or jumps in the level of the resulting impulse response. These artefacts can be easily detected by the human ear and significantly deteriorate the overall perceived quality.

IMPLEMENTATION OF PERCEPTUAL APPROACH

MPEG-4 Audio provides the *Perceptual Approach* as an alternative to room acoustic rendering of audio based on the geometry and material characteristics of the virtual scene. Instead, here a number of parameters (the so-called perceptual parameters) are provided with which the characteristics of a filter network representing an artificial room impulse response (amplitude-, frequency-, and time-wise) can be influenced. MPEG-4 Advanced Audio BIFS specifies these parameters, and the standard provides detailed infos on which the implementation presented here is based. Fig. 3 shows an intermediate transformation step of the schematic view of the model as described in the standard.

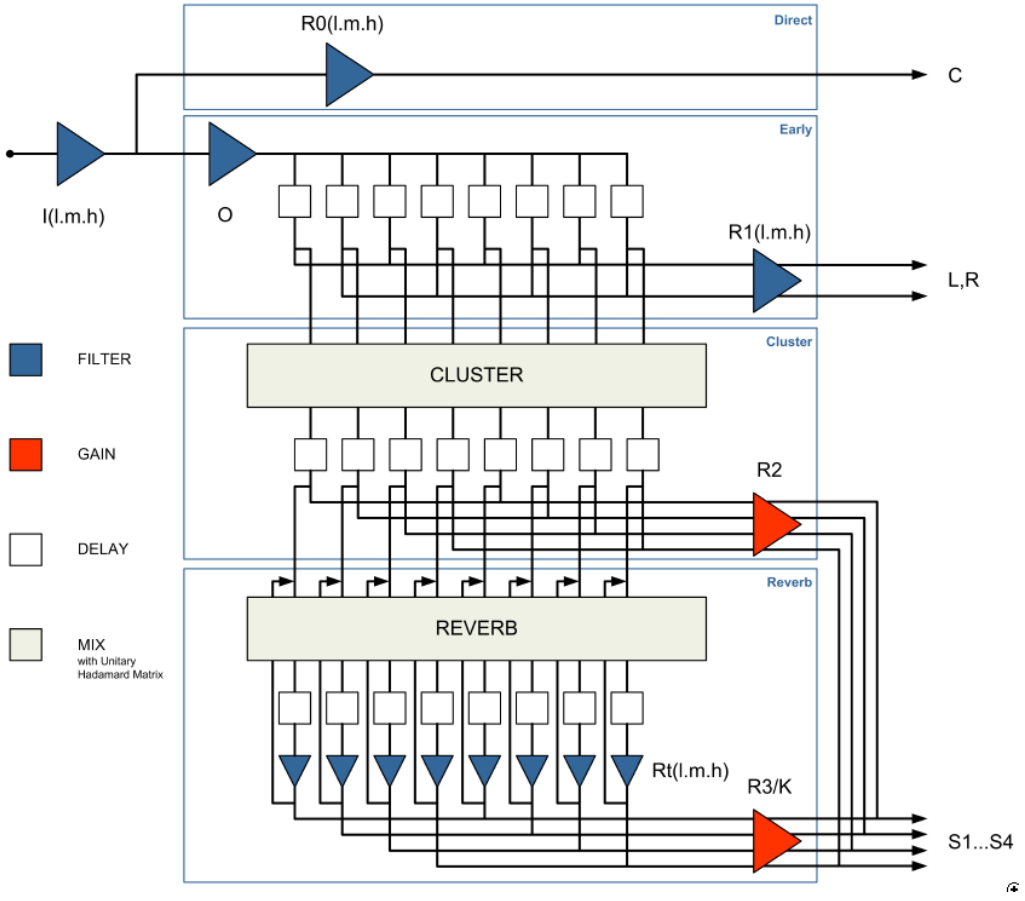

Fig. 3: Intermediate transformation step of the IRCAM-model. For details compare [5].

The *TangaPerceptualSource* Util represents the top layer of the implementation of the Perceptual Approach in the TANGA engine. It is therefore the interface for other classes which need to make use of the *Perceptual Approach* model. It hides all of its signal and data processing and can be regarded as a summarization or composition of Components to form a complex signal processing unit. An instantiation can only happen if the loudspeaker setup, the number of loudspeaker channels to be used, the perceptual parameters themselves, the distance and position of the sound source, the workchannelorder[1] as well as a unique ID are passed as parameters. Fig. 4 shows the client interface of the *TangaPerceptualSource* Util. This interface is a typical example for other Utils used within the TANGA engine.

[1] The generic aspect of this implementation of the IRCAM-model consists in the scalability of the number of processing channels, here called workchannels. The number of workchannels to be used is defined as: workchannels = workchannelorder 4.

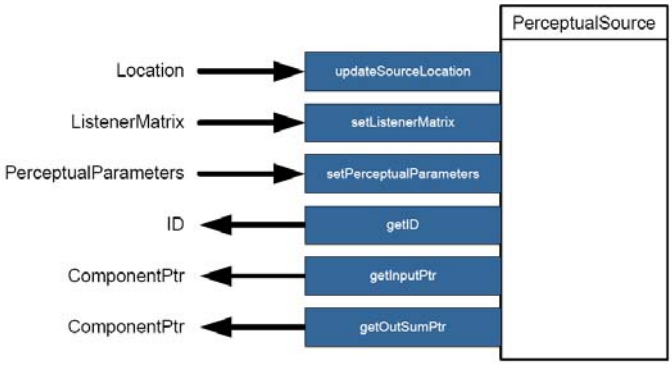

Fig. 4: The client interface of the TangaPerceptualSource Util.

In fig. 5 the control operations necessary in the *TangaPerceptualSource* Util can be seen. For the meaning of the data designators (low level parameters) see [5]. All parameters either are passed from the scene description directly (perceptual parameters) or are derived from them in the *calculateHighLowMapping* method.

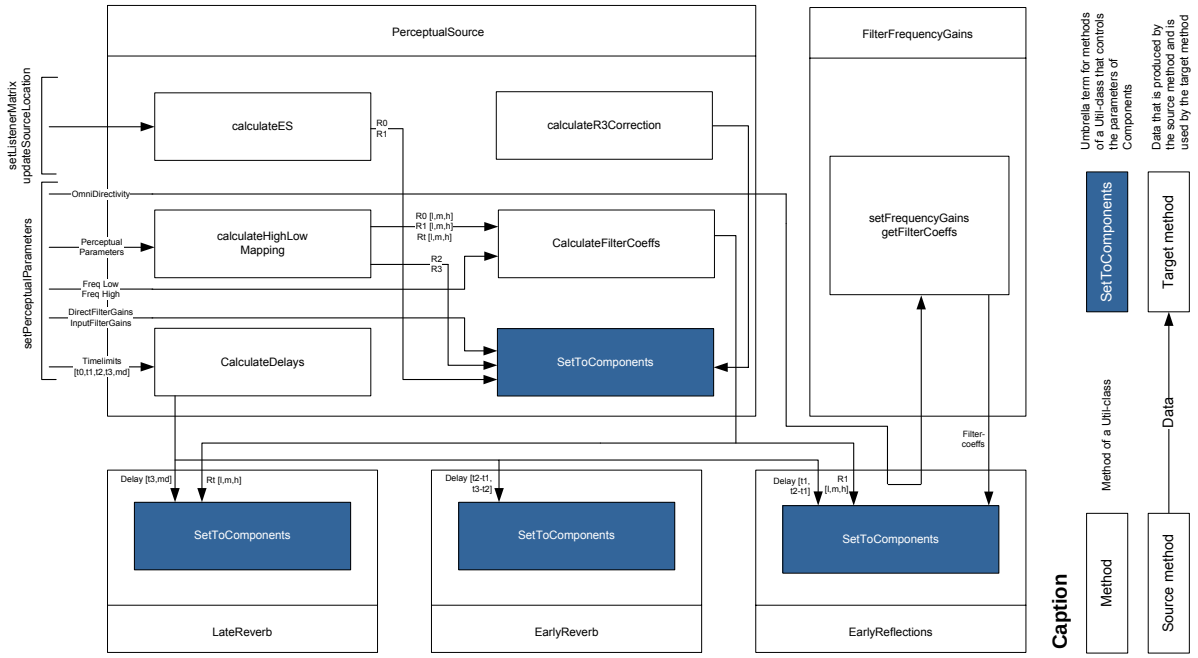

Fig. 5: The control operations in the TangaPerceptualSource Util.

Fig. 6 shows the Component Graph of the TangaPerceptualSource Util which encapsulates all Components and Utils needed to apply the Perceptual Approach to an audio signal. As can be seen, it consists of a number of other Utils (EarlyReflections, EarlyReverb, LateReverb) and Components (TangaFilter, TangaVBAP3, TangaMatrix, TangaMix). This is the usual way the auditory functionality is constructed in the TANGA

engine: according to the (A)AudioBIFS[2] nodes contained in the scene description, the necessary auditory functionality is composed from smaller elements by connecting Components and Utils. Functional groups can be organized in Utils, which themselves can be further grouped to form part of other Utils.

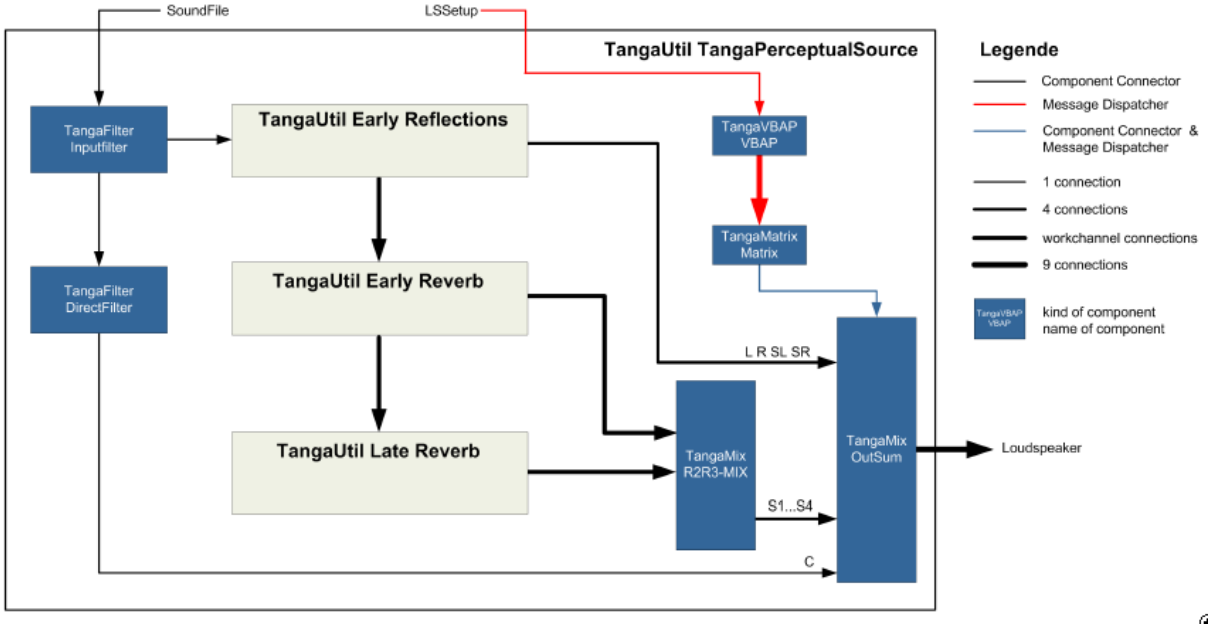

Fig. 6: Component Graph of the TangaPerceptualSource Util.

The MPEG-4 Audio *perceptual approach* can also be computed with varying complexity in the TANGA engine. The implementation exceeds the standard by far. This is necessary to evaluate the cost-value ratio of the algorithm with respect to the overall perceived quality. A detailed study of internal complexity of the algorithm used can be found in [6]

IMPLEMENTATION OF ACOUSTIC OBSTRUCTION

When working in a dynamic bimodal (audio visual) environment, acoustic and visual objects need to be synchronous and concordant in order to provide some degree of scene realism. MPEG-4 provides means for making sure that auditive and visual content are rendered “in sync”. Whenever self-movement of the user through the scene is allowed, AV synchrony and concordance is even more important for the scene realism than high quality rendering [7].

Therefore we have implemented the detection and simulation of acoustic obstruction in the I3D. It is based on visual obstruction, i.e. whenever a sound source is not visible

[2] (Advanced) Audio BInary Format for Scene description

because of some obstructing object blocking the line of sight between user and sound source, the acoustic properties of the obstructing object are checked. If the obstructing object is acoustically relevant, then acoustic obstruction is simulated.

The detection of obstruction has to be performed in real time. Because there are usually a large number of objects which could potentially obstruct the sound source, a very fast algorithm needs to be applied. In the I3D, we use an intersection test with a Bounding Sphere[3] around the potentially obstructing object for higher performance. By this, we do not need to check all polygons of the object for obstruction, but the algorithm has only to be performed once on the complete BSphere in each render pass. Fig. 7 shows a schematic view of the intersection test performed with a Bounding Sphere around a potentially obstructing object.

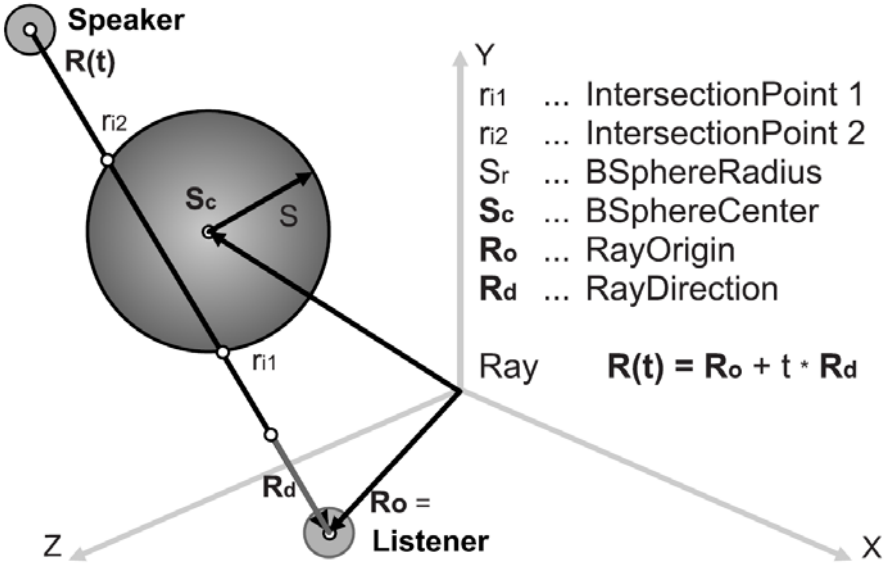

Fig. 7: Intersection test with a Bounding Sphere [8].

The difference between intersection points n_1 and n_2 versus the BSphere radius S_r indicates the degree of obstruction. Obstruction is maximal when

$$|n_1 - n_2| = 2S_r \quad \textbf{(1)}$$

and minimal when

$$n_1 = n_2 \quad \textbf{(2)}$$

A subjective assessment among 21 test subjects was performed, evaluating the optimum transition curve between the two conditions ”obstructed” and ”not obstructed” sound source. Subjects were asked to rate different transition curves of exponential,

[3] A Bounding Sphere or BSphere is the smallest possible sphere that encloses an object completely.

linear and logarithmic slope generated according to equ. 3 by varying x between x = 1/5 and x = 5. *transfunc* represents the acoustic transparency factor of the obstructing object, whereas the degree of obstruction is equal to 1 − a/S_r.

$$gain = \left(\left(\frac{a}{S_r} \right)^x \cdot (1 - transfunc) \right) + transfunc \qquad \textbf{(3)}$$

For the given test setup (circular 8-channel loudspeaker setup, projecting screen of 2.7m of width, see fig. 8) the differences between the transition curves offered to the subjects were too subtle to be relevant. Therefore the least computationally demanding algorithm, a linear transition (x = 1) between the two conditions, is sufficient. The full details of the assessment including the complete statistical analysis can be found in [9].

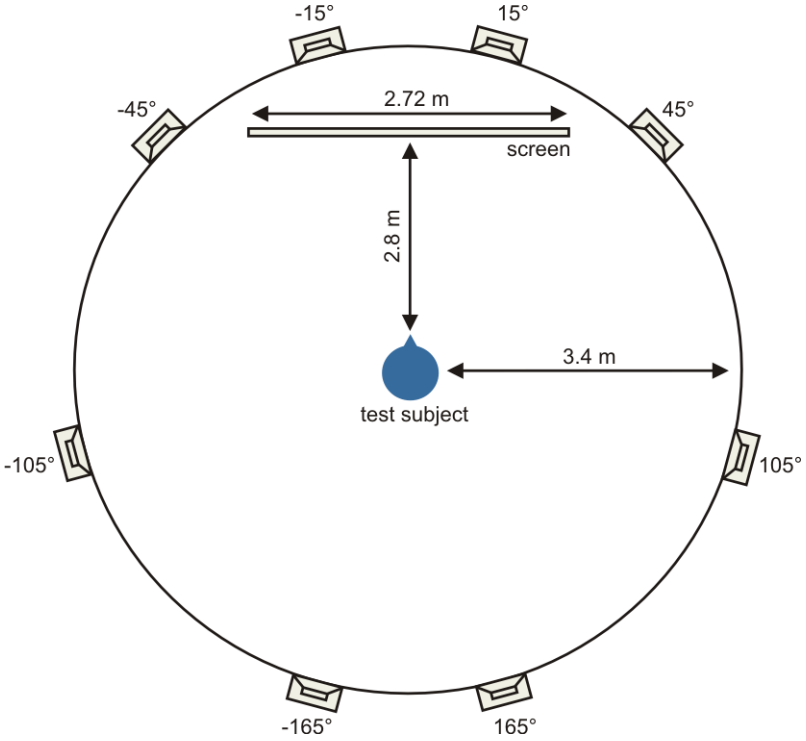

Fig. 8: Loudspeaker and projecting screen setup used for the subjective assessment of acoustic obstruction.

MIDI INTEGRATION

The I3D can be controlled by a number of input devices. These are keyboard, mouse and the MIDI interface. The MIDI input/output is used to allow test subjects to control the assessment and to interact with the scenes to be assessed. For this we designed a control desk which displays status information coming from the I3D player. The motorized faders are used by the test subjects to enter their ratings, whereas the buttons with status indicator LEDs are used to select the items to be rendered by the I3D.

Fig. 9 shows the MIDI communication structure used in the assessments: PC1 is running

the I3D, PC2 is running the data logging tool SALT6, and the test subject is handling the control desk. Upon completion of a test session, SALT writes a log-file which contains individual-related test subject data, semantic designators of trials and stimuli as well as a matrix with trial IDs, item IDs and item ratings. The order in which the trials / items appear in the matrix corresponds to the (stochastic) order in which these were presented to the test subject. After an assessment is completed, all matrices can be combined and saved as one MATLAB workspace for further statistical analysis.

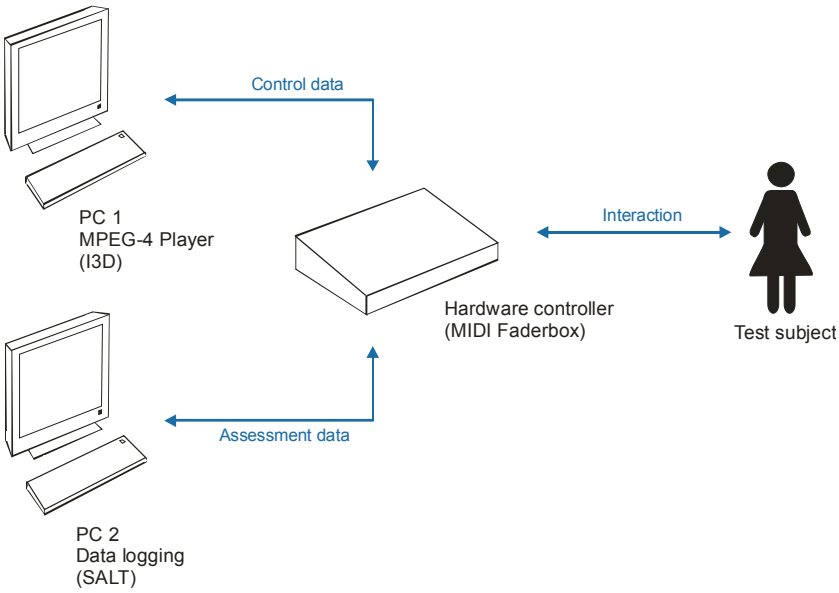

Fig. 9: MIDI communication between the control desk, I3D (PC1) and logging tool (PC2).

OTHER FEATURES

Currently, the TANGA engine comprises a number of Components which either represent a subset of the MPEG-4 Audio nodes, or enhance the functionality by implementing important features especially needed for increased AV quality. These are TangaDelay, TangaDelayLine, TangaDiffuseReverb, TangaDirectivity, TangaFilter, TangaFilterBlender, TangaLSSetup, TangaMatrix, TangaMix, TangaObstruction and TangaSum. The number of Components available is increasing continuously as the engine evolves.

From these Components and combinations of Components (e.g. encapsulated in Utils), a number of MPEG-4 Audio features can be implemented directly. Among these are e.g. the *DirectiveSound* node, which allows defining the frequency-dependent radiation pattern of directional sound sources.

SUMMARY AND OUTLOOK

We have implemented the two different methods for creating room acoustic effects specified in the MPEG-4 standard ISO/IEC 14496-1:2004, the Physical Approach as well as the Perceptual Approach. Furthermore, the I3D is capable of handling one of the most important audio-visual effects, obstruction of sound sources, satisfactorily. The implemented features along with the integration of MIDI communication capabilities can be used to perform audio-visual subjective assessments, which are necessary to further improve the performance and quality of the system.

The basic structure of the TANGA engine is very flexible and allows arbitrarily complex algorithms to be included into the I3D. It is therefore a very elegant basis for audio-visual subjective assessments in general: not only for improving performance of the system itself, but also to gain knowledge about the cognitive processes behind human bimodal (audio visual) perception.

Future work will include the improvement of the detection of obstruction in terms of precision and performance, as well as the integration of two-dimensional sound sources (MPEG-4 AudioBIFS v3 draft *WideSound)* as an addition to the point sound sources implemented by now. These may play an important role in the perception of audio visual scenes in interactive AV environments.

ACKNOWLEDGMENT

This work is supported by the EC within FP6 under Grant 511568 with the acronym ’3DTV’.

References:

[1] Drumm, H.; Kühhirt, U.; Rittermann, M.; Reiter, U.: “Application Systems for MPEG-4”, IEEE/ISCE’02, International Symposium on Consumer Electronics, Erfurt, Germany, 23rd-26th September 2002

[2] Reiter, U.: “On the Need for a Salience Model for Bimodal Perception in Interactive Applications”, IEEE/ISCE'03, International Symposium on Consumer Electronics, Sydney, Australia, 3rd-5th December 2003

[3] Reiter, U., Köhler, T.: “Criteria for the Subjective Assessment of Bimodal Perception in Interactive AV Application Systems”, IEEE/ISCE'05, International Symposium on Consumer Electronics, Macau, SAR/China, June 2005, ISBN 0-7803-8920-4

[4] Dantele, A., Reiter, U.: “Description of audiovisual virtual 3D scenes: MPEG-4 perceptual parameters in the auditory domain”, IEEE/ISCE'04, International Symposium on Consumer Electronics, Reading, UK, September 2004, ISBN 0-7803-8526-8

[5] ISO/IEC 14496-1:2004, Information technology – Coding of audio-visual objects, part 11.1, MPEG-4 Systems Node Semantics, 2004.

[6] Reiter, U.; Partzsch, A.; Weitzel, M.: “Modifications of the MPEG-4 AABIFS Perceptual Approach: Assessed for the Use with Interactive Audio-Visual Application Systems”, Proceedings of the AES 28th International Conference, 2006 June 30 to July 2, Pitea, Sweden

[7] S. Zielinski, F. Rumsey, S. Bech, B. de Bruyn, and R. Kassier, “Computer Games and Multichannel Audio Quality - The Effect of Division of Attention Between Auditory and Visual Modalities,” in Proceedings of the AES 24th International Conference on Multichannel Audio, Banff, Alberta, Canada, 2003, pp. 85–93.

[8] Eric Haines, An Introduction To Ray Tracing, Academic Press Limited, 1989.

[9] Steglich, B; Reiter, U.: “Sound Source Obstruction in an Interactive 3DimensionalMPEG-4 Environment,” in Proceedings of the AES 120th Convention, Paris, France, 2006.

[10] W. G. Gardner, “A Realtime Multichannel Room Simulator”, J. Acoust. Soc. Am., 92(4), pp 2395, and presented at the 124th Meeting of the Acoustical Society of America, New Orleans, USA, November 1992

[11] PortAudio, an Open-Source Cross-Platform Audio API, http://www.portaudio.com/

Author:
Dipl.-Ing. Ulrich Reiter
Institut für Medientechnik
Technische Universität Ilmenau, Helmholtzplatz 2
98693 Ilmenau, Germany
Phone: +49 – 3677 – 69 2671
Fax: +49 – 3677 – 69 1255
E-mail: ulrich.reiter@tu-ilmenau.de