Evaluation of voxel-based rendering of high resolution surface descriptions

Dorte Hammershøi, Søren Krarup Olesen, Miloš Marković[a]
Acoustics, Department of Electronic Systems, Aalborg University, Fredrik Bajers Vej 7 B5, 9220 Aalborg Ø, Denmark

Summary
The auditory rendering in a VR system either entail real, acoustical measurements of the room, or – more commonly – a model of the room rendered. The models are usually based on relatively coarse approximations of the geometry, and the calculated room impulse responses deviate from measured impulse responses by having more "distinct" representations of the individual reflections. When rendered audible, such descriptions will lead to a sound quality that can best be described as "canned". For the rendering of real rooms, as e.g. in "teletransporting", this problem may be addressed by using high resolution scans of the room surfaces as basis for the room models. In the present work this approach is evaluated with a voxel-based method and compared to measured impulse responses. The results are compared objectively by visual inspection of the impulse responses, and by comparison of statistical parameters calculated from measured and modelled data. The work is ongoing.

PACS no. 43.55.+p, 43.60.+d

[a]now at Huawei Technologies Duesseldorf GmbH, ERC Riesstr. 25, 80993 Munich, Germany

1. Introduction

In telepresence and virtual reality (VR) it is desirable to have a high degree of perceptual immersion, so that the human experiencing the virtual world can take advantage of all natural hearing abilities, e.g. the cocktail-party processor. The data (the signal) itself, for example speech, must not only be constructed or transferred in a way that makes the speech understandable. It should also be transferred in a way that makes the speech convincing to the extent that listener acts as if the speaker was physically present [1] in the same environment. This is part of the vision for the EU telepresence project BEAMING[1], where one person (the *Visitor*) can experience a remote scene inhabited with *Locals*.

A general audio scenario consists of individual acoustical events (a person speaks, a signal turns on, etc.), each associated with (sometimes moving) objects, plus an acoustical contribution from the boundaries of the room the objects are within.

Network-oriented speech communication systems (e.g. Skype) often rely on close-up microphones or, at least, microphones placed in close proximity to the speakers so a good signal-to-noise ratio is obtained, and a high speech intelligibility is achieved. Such microphone set-ups do pick up portions of the room acoustics, but unintentionally and with no prospect for authentic rendering remote. For the remote rendering, it is preferable to have the individual audio streams clean from the room acoustics so that this may later be added in a way that puts the speaker and listener in the right positions relative to each other in the shared physical/virtual environment.

The BEAMING project allows for exactly that [2]. A period assigned for adjustment is available (approximately 30 mins) before the communication session takes place and can be used for either impulse response measurements or other methods that can capture the room acoustically. Alternatively, the period can be spent on scanning the room geometrically, and record what we shall denote point-clouds in the following.

1.1. The point-cloud

A point-cloud is a large set of 3D-points which describes the surrounding geometry. The larger the set, and the denser the points, the better is the description of the geometry which in this case is a room. We have used the Kinect for Xbox 360 to acquire the geometrical data. The Kinect contains a depth (range detection) camera that uses infra-red light patterns, and the depth map is associated with a colour frame.

[1] http://beaming-eu.org/

The camera can e.g. be placed in the middle of the room and automatically rotated on a stand.

The outcome depends on the shape and size of the room. The precision decreases at longer distances to boundaries and if the room contains obstacles seen from the point of the camera, "holes" and shadowed areas might appear. Changing the position of the camera can remove such areas and methods exist that combine point-clouds from different positions in order to form a more complete map of the room.

Once the room has been mapped the point-cloud undergoes a number of numerical transformations, here using the open-source Point Cloud Library (PCL), after which the data is passed to a room simulation process [3, 4].

1.2. Room simulation

Acoustic objects move in all real world scenarios of interest, thus the contribution from the room changes continuously, too. It is in principle possible to measure impulse responses from and to all possible positions in a room, but it takes time, require storage space at the rendering engine, and needs specialized equipment at the remote scene of interest. An alternative is to initially capture sufficient information of the room, so that it can be modelled, and estimates of the sound transmissions computed on the fly.

When the point-cloud and hence a geometrical description of the room is handy there is in principle several ways to go. Utilizing PCL the points can be used to construct a triangular mesh of surfaces, which paves the way for classic image and ray-tracing methods, as used in commercial analysis software, e.g. ODEON [2], CATT[3], or EASE[4].

Another more direct approach is to take advantage of the intrinsic granularity of the original points. This granularity is used to construct a grid of voxels and subsequently apply voxel based "discrete raytracing" [5]. This has two advantages; 1) converting points into voxels is a fast process and 2) voxel based room simulation is itself a relative fast process.

The output of the room simulation is an impulse response. In order to confirm that the point-cloud based, discretized room simulation process is capable of producing a correct impulse response, it is compared to an impulse response measured (not modelled in any way) in the real room, see Figure 1.

As can be seen on Figure 1, the dashed loudspeaker and microphone (both virtual) positions can be moved arbitrarily, since that only affects the positional information to the room simulation program. The scan of the geometry of the rooms boundaries (all fixed objects) thus only needs to be carried out once, e.g. during the initialization phase prior to activating the BEAMING system.

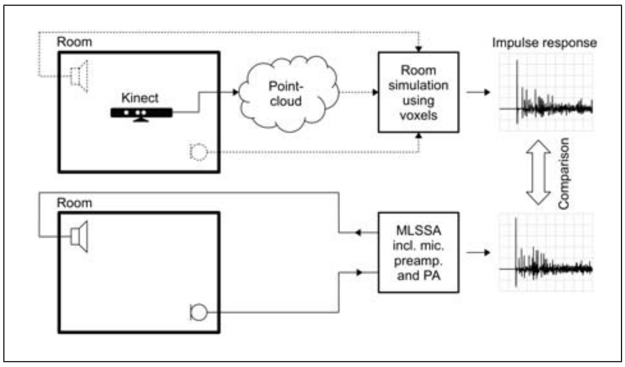

Figure 1. The point-cloud captured by the kinect is used as input for a room simulation program, which can compute impulse responses for any position of source and receiver in the measured and modelled environment (top). Another impulse response is measured (by a MLSSA system [6]) directly in the room for which the geometry has been measured (bottom).

1.3. Purpose of investigation

The purpose of the present investigation is to compare impulse responses rendered using the point-cloud methods with measured impulse responses for the same room. The present paper includes a presentation of the room impulse response measurements, incl. estimation of source and receiver positions to be used in the room simulation program. A tentative report on the present study was given in [7].

2. Methods

2.1. Room

Impulse response measurements were made in a rectangular room which has been scanned earlier [3, 4]. The room has relatively hard walls and ceiling, whereas the floor is carpeted. One side and one end of the room is fairly complicated geometrically with windows, cable panels, and a door. The opposing side and the other end wall is perfectly straight and geometrically simple, see Figure 2.1. Measurements were made in different source-receiver positions incl. configurations close to the complicated wall, close to the simple wall, close to one uncomplicated corner, close to a complicated corner, and not close to any of the walls.

2.2. Source-receiver configurations

A total of 11 measurements were performed, varying both the heights of microphone and loudspeaker, as well as the distances to the walls (and ceiling). Positions were selected arbitrarily to cover a fair range of options in the room, incl. positions close to the closest boundaries, and positions far from boundaries, and

[2] `htth://www.odeon.dk/`

[3] `htth://www.catt.se/`

[4] `http://ease.afmg.eu/`

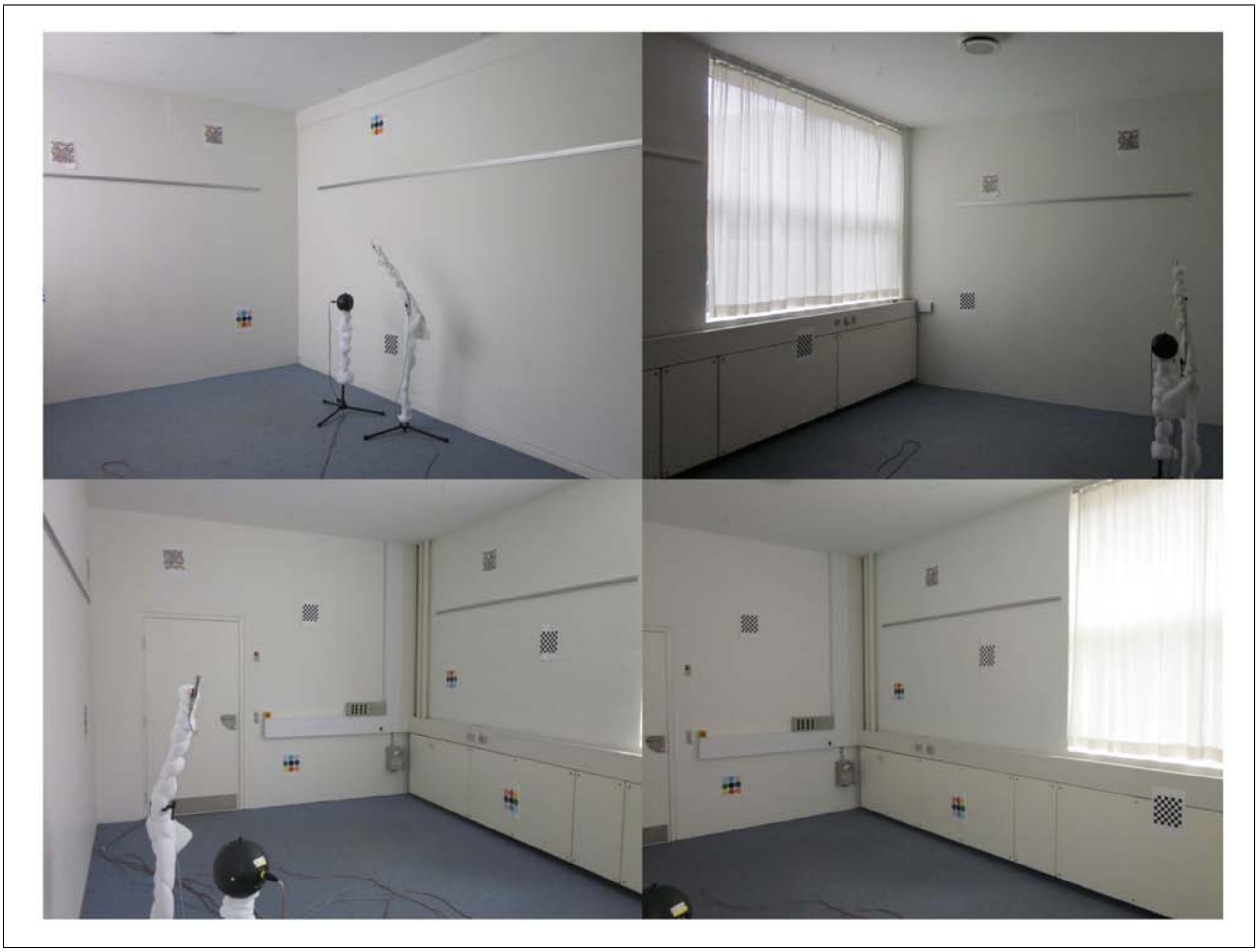

Figure 2. Pictures from the room measured.

also configurations with source and receiver close or far from each other. No attempt was made to measure in positions were room modes would have nodes or anti-nodes. The distances were measured with an infra-red meter with an accuracy of approximately 1-2 mm. The primary source of inaccuracy in the distance measurements was the operator.

The measurement positions are listed in Tables I and II.

Another two measurements were performed after the first and second measurement with the autorange option of the MLSSA system disabled, and the input to the power amplifier short-circuited. These measurements are plotted in Figure 2.4 for evaluation of the signal to noise ratio.

2.3. Measurement instrumentation

The measurements were made using maximum length sequence methods as implemented in the MLSSA system (DRA Laboratories) [6]. A 16^{th} order MLS sequence was used with a 50 kHz sampling rate resulting in impulse responses of 1.31 s length, which is sufficient to avoid time aliasing for the given room. The built-in Chebyshev anti-aliasing filter was used with a cut-off frequency at 20 kHz.

Table I. Measured position of source (center of ball loudspeaker).

Number	Δx [m]	Δy [m]	Δz [m]
1	1.559	0.815	0.962
2	3.255	1.545	0.962
3	2.825	0.649	1.434
4	2.825	0.649	1.434
5	4.491	1.773	1.660
6	4.924	2.120	1.665
7	4.435	3.042	1.665
8	3.579	1.511	1.665
9	3.911	2.994	1.665
10	2.591	1.741	1.425
11	5.202	2.017	0.972

Table II. Measured position of receiver (center of microphone diaphragm).

Number	Δx [m]	Δy [m]	Δz [m]
1	0.459	0.448	1.221
2	1.050	2.019	1.221
3	1.050	2.049	1.221
4	1.686	1.101	1.221
5	1.778	1.308	0.463
6	1.708	1.697	1.695
7	1.708	1.697	1.695
8	2.462	1.597	1.695
9	1.949	2.972	1.695
10	2.644	0.347	1.284
11	5.407	2.601	1.048

The sound stimulus was played back by an in-house produced ball loudspeaker, similar to loudspeakers used in previous investigations, see e.g. [8] for an example of the frequency response. This loudspeaker type is not perfectly omnidirectional, but has a radiation pattern much like that of the human [9]. The loudspeaker was always positioned to point directly towards the measuring microphone.

A B&K type 4166 condenser microphone with a nominal sensitivity of 47.3 mV/Pa was used. The microphone was always positioned so that the direct sound would impinge perpendicularly to the microphone diaphragm. A B&K 2639 pre-amplifier was used, and connected to a B&K 2636 measurement amplifier with a 40 dB amplification. The 5 V full scale deflection output was used, adding extra 13.98 dB amplification (a gain factor of 5).

The main source of noise was the ventilation system. The room was well ventilated at the beginning of the measurements, and since the room was not under influence of any heat sources of significance, ventilation was turned off during measurements. A signal to noise ratio of approximately 20 dB (and better) was obtained for frequencies above approximately 100 Hz, see Figure 2.4.

2.4. Example of measurements

The data are converted into Matlab format using inhouse developed interface software, which correct the amplitude of the impulse response for the stimulus amplitude (known MLSSA flaw). Figure 3 shows an example of one of the measurements.

Figure 3 shows excerpts of the 1.31 s (65535 point) long impulse response measured at source-receiver configuration No. 1. The left panel (first 200 ms) shows the typical, distinct characteristics of a room impulse response with a strong direct sound and early reflections, followed by a decaying, more diffuse tail. From the right panel (first 20 ms) it is possible to identify the individual early reflections.

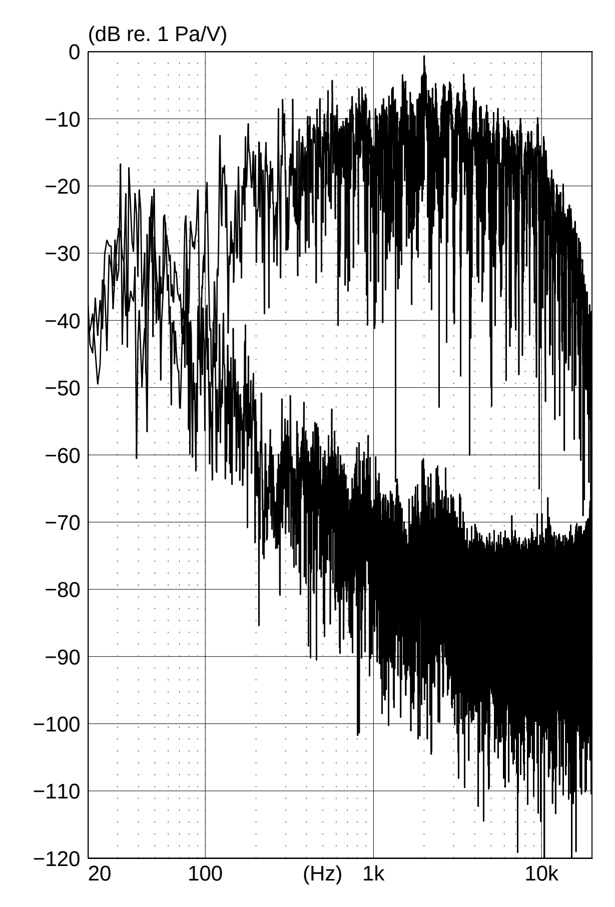

Figure 4. Frequency response of a typical measurement (source-receiver configuration 1), incl. frequency spectrum of estimated noise floor.

The comparison with simulations shall evaluate both the accuracy of arrival time as well as the nature of the distinct reflections, and similarity of typical room acoustical characteristics (reverberation times, early-late ratios, etc.).

Figure 2.4 (top) shows the frequency response of the measurement for source-receiver configuration No. 1. All 65535 measured points are included in the Fourier transform. The frequency response shows the distinct room response patterns with narrow dips from the room reflections.

The overall shape of the frequency response is a soft band-pass characteristic. This is the characteristics of the source (the loudspeaker), and not the room. No attempt is made to compensate for this, since the SNR is too poor outside the bandpass range to provide any information of validity.

The second curve shown in Figure 2.4 (bottom) is an estimation of the noise floor. This is a Fourier transform of the measurement made with all settings similar, but with the auto-range option disabled and with the input of the power amplifier for the loud-

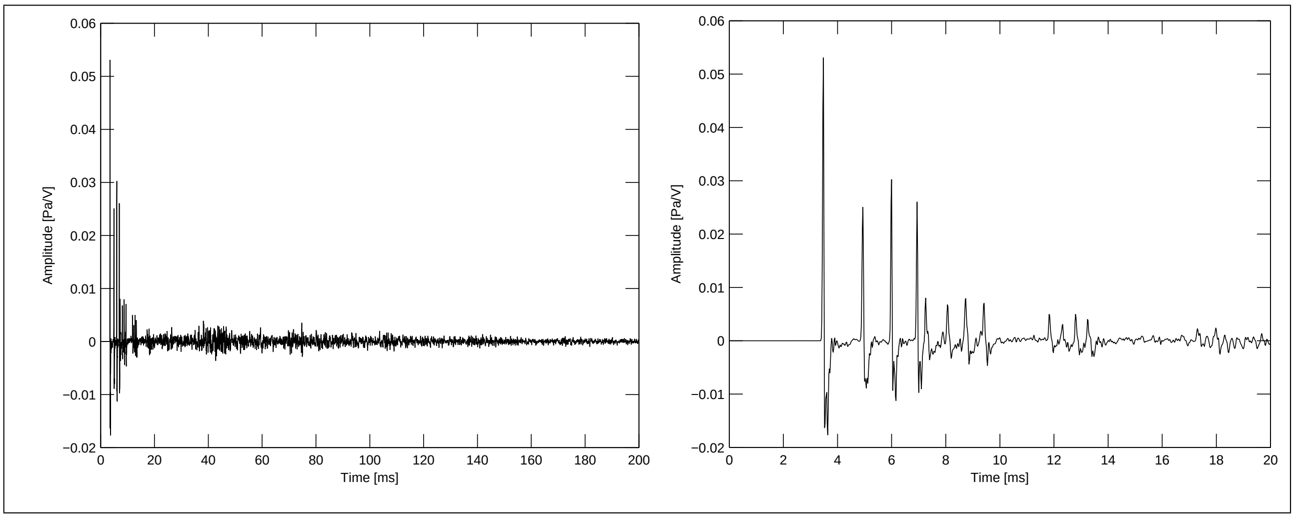

Figure 3. Excerpts of raw impulse response (No. 1). Left panel shows the first 200 ms (samples 1:9999), and right panel show the first 20 ms (samples 1:999).

speaker short-circuited. A signal to noise ratio better than 20 dB is obtained for frequencies above approximately 100 Hz up to approximately 20 kHz. This is considered sufficient for the present purpose.

2.5. Estimation of source position

The position data are made with reference to the center of the ball in which the Vifa unit is mounted. This is an arbitrary point with respect to acoustic radiation, and an estimate of the acoustic center is required. This is not trivial for any source, but in the present case estimates can be based on the measurements at hand.

The general strategy for estimating the acoustic center of the source is to determine the true arrival times of the direct sounds (t_{true}) in all measurements, compare these to the arrival times that can be estimated from the physical positions measured (t_{est}), and use the systematic difference in $\Delta t = t_{true} - t_{est}$ to compute position values for the source that better represent the true acoustic source.

t_{est} is computed like this for each source-receiver configuration:

$$t_{est} = c \cdot f_s \cdot \|\vec{r} - \vec{s}\| \quad (1)$$

where c is the speed of sound (assumed to be 344 m/s at 20°C), f_s the sampling frequency (50 kHz), and $\vec{r}$ and $\vec{s}$ represent vectors for the receiver (the center of the microphone diaphragm) and the source (the center of the ball loudspeaker), i.e. the measured positions, see Table I.

t_{true} is determined in two ways, one using visual inspection, t_{vis}, and one based on an identification of the first point exceeding 5% of the absolute maximum value of the impulse response, t_{max}.

When zooming in on the first samples of the impulse (see Figure 5), it can be seen that t_{est} overshoots the distance by a few samples. This is the general picture, and in agreement with expectations.

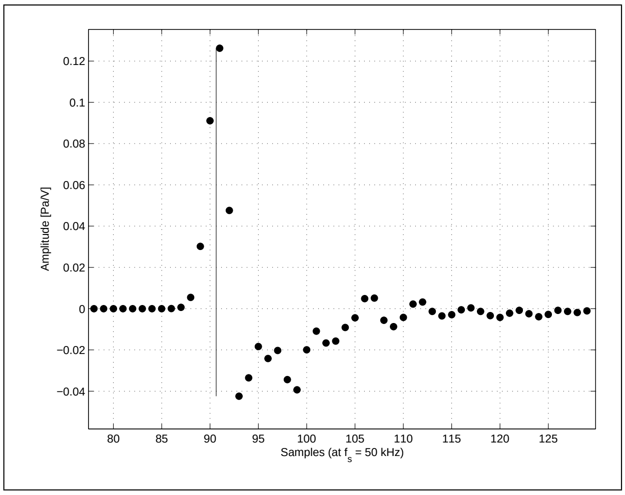

Figure 5. Close-up look at the initial part of a measured impulse response (No. 11). The black points represent the measured samples, and the vertical line represents the arrival time that would correspond to the measured position values for the source (the center of the ball loudspeaker).

Two of the measured impulse responses revealed a surprisingly low and a surprisingly high deviation in the arrival time estimates.

One of the two impulse responses that show a difference is that of source-receiver configuration No. 1, the first impulse response measured. The difference for this impulse response is close to zero (less than one sample) and positive, which means that the position measured corresponds to a point some millimetres in front of the acoustic center. It is assumed that the distance for this measurement was by mistake measured from the front of the speaker diaphragm, and not from the center of the ball loudspeaker.

The other impulse response that deviates is that for source-receiver configuration No. 3. The difference between estimated travel time and apparent travel time is almost twice as large as the general average. It is assumed that a measurement error occurred, which could be on either the source or receiver position.

Both measurements (Nos. 1 and 3) are excluded from the further analysis.

The remaining nine measurements showed systematically shorter travelling times for the sound wave than the estimated travelling time referring to the loudspeaker center. The average Δt based on t_{vis} visual inspection equals 3.2 samples. The average Δt based on t_{max} equals 1.7 samples. A conservative estimate based on the average of the two (equal to 2.4 samples, and corresponding to a distance of $2.4/50000 \cdot 344 = 0.0165$ m) is used as a correction value in the estimation of the true source center position. The modified source positions, s_m are determined as:

$$s_m = \frac{\vec{s} - \vec{r}}{\|\vec{s} - \vec{r}\|} \cdot \left(\|\vec{s} - \vec{r}\| - \frac{\Delta t}{f_s \cdot c} \right) + \vec{r} \tag{2}$$

A final re-calculation of source and receiver positions is required to place them in the same coordinate system as was used when scanning the room. The kinect device was positioned in the center of the room horizontally, and at a height of 1.13 m. The coordinate system was standard right-hand turned with the position of the kinect device in origo. The positions for the measurement coordinate system is transferred to the coordinate system for the point cloud by the following:

$$r_k(x, y, z) = \left(\frac{w}{2} - \Delta y, \frac{l}{2} - \Delta x, \Delta z - 1.13 \right) \tag{3}$$

where $l = 5.70$ m is the length of the room, and $w = 3.55$ m is the width. The resulting coordinates are listed in Tables III and IV.

3. Results

Measurements with high signal to noise ratio have been obtained for the room, for which a geometrical scan exists. The high SNR makes it possible to make precise comparisons to the room simulations, which–in principle–have an infinite SNR. The measurements provide the ability to clearly identify arrival time and energy of low-order reflections, and the room simulations must match those quantities.

4. Discussion

The system topology of the BEAMING audio system is based on binaural methods [10, 11, 12, 13]. The impulse responses measured in the present investigation relate to one single point in space, and can't be used in listening experiments. The alternative would have been to measure the room impulse response with a manikin [14] (or a human head [15]), and use the head-related transfer functions of the manikin for binaural auralization [16]. A choice was made against this, because it would be easier to identify the individual reflections in the measurements, if they weren't convolved with the characteristics of the manikin.

Table III. Position of source (acoustic center) in point-cloud coordinate system.

Impulse no	x	y	z
2	0.2266	-0.3890	-0.1661
4	1.1200	0.0401	0.3012
5	0.0046	-1.6261	0.5234
6	-0.3428	-2.0576	0.5352
7	-1.2597	-1.5702	0.5352
8	0.2627	-0.7125	0.5354
9	-1.2188	-1.0445	0.5353
10	0.0504	0.2584	0.2933
11	-0.2575	-2.3574	-0.1560

Table IV. Position of receiver (center of microphone diaphragm) in point-cloud coordinate system.

Impulse no	x	y	z
2	-0.2440	1.8000	0.0910
4	0.6740	1.1640	0.0910
5	0.4670	1.0720	-0.6670
6	0.0780	1.1420	0.5650
7	0.0780	1.1420	0.5650
8	0.1780	0.3880	0.5650
9	-1.1970	0.9010	0.5650
10	1.4280	0.2060	0.1540
11	-0.8260	-2.5570	-0.0820

5. CONCLUSIONS

The simulations are presently ongoing.

Acknowledgement

This project has been funded by the European Commission through grant FP7 248620 – BEAMING. The authors would like to acknowledge the consortium partners, in particular partners from Department of Computer Science, University College London and in particular Fabrizio Pece, also UCL, for his help in room geometry acquisition using depth cameras.

References

[1] Y. A. Huang, J. Chen, J. Benesty: Immersive audio schemes. *IEEE Signal Processing Magazine* **28**20–32 (2011).

[2] S. K. Olesen, M. Marković, E. Madsen, P. F. Hoffmann, D. Hammershøi: 3D sound in the telepresence project BEAMING. *Proc. BNAM2012, Odense, Denmark*, pp. 1–4 (2012).

[3] M. Marković, S. K. Olesen, D. Hammershøi: Three-dimensional point-cloud room model for room acoustics simulations. *Proc. 21th ICA 2013*, Montreal, Canada, pp. 1–7 (2013).

[4] M. Marković, S. K. Olesen, D. Hammershøi: Room acoustics modeling using a point-cloud representation of the room geometry. *Proc. ISRA 2013*, Toronto, Canada, pp. 1–6 (2013).

[5] S. K. Olesen: An integer based ray-tracing algorithm. *Proc. 100th Audio Eng. Soc. Conv.*, Copenhagen, Denmark, preprin 4227:1–6 (1996).

[6] D. D. Rife, J. Vanderkooy: Transfer-function measurement with maximum-length sequences. *J. Audio Eng. Soc.* **37**(6) 419–444 (1989)

[7] S. K. Olesen, M. Marković, D. Hammershøi: Fast rendering of scanned room geometries. *Proc. BNAM2014*, Tallinn, Estonia, pp. 1–7 (2014).

[8] H. Møller, M. F. Sørensen, D. Hammershøi, C. B. Jensen: Head-related transfer functions of human subjects. *J. Audio Eng. Soc.* **43**(5) 300–321 (1995).

[9] S. H. Nielsen: Distance perception in hearing. Ph.D. thesis, Aalborg University, 1991.

[10] J. Blauert: Spatial Hearing. *The MIT Press*, 1997.

[11] H. Møller: The fundamental of Binaural Technology. *Appl. Acoust.* **36**(3–4)171–218 (1992).

[12] J. Blauert, H. Lehnert, J. Sahrhage, H. Strauss: An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation. *Acustica* **86**(1)94–102 (2000).

[13] D. Hammershøi, H. Møller: Binaural technique – Basic methods for recording, synthesis, and reproduction. Chapter 9 in: *Communication Acoustics*, Ed. J. Blauert, Springer Verlag, 2005.

[14] H. Møller, D. Hammershøi, C. B. Jensen, M. F. Sørensen:Evaluation of artificial heads in listening tests. *J. Audio Eng. Soc.* **47**(3)83–100 (1999).

[15] H. Møller, M. F. Sørensen, C. B. Jensen, D. Hammershøi: Binaural technique: Do we need individual recordings? *J. Audio Eng. Soc.* **44**(6)451–469 (1996).

[16] H. Møller: Interfacing room simulation programs and auralisation systems. *Appl. Acoust.* **38**(3–4)333–347 (1993).