Audio Engineering Society

Convention Paper

Presented at the 125th Convention
2008 October 2–5 San Francisco, CA, USA

A platform for audiovisual telepresence using model- and data-based wave-field synthesis

Gregor Heinrich[1,2], Michael Leitner[1], Christoph Jung,[1], Fabian Logemann[1], and Volker Hahn[1]

[1]*Fraunhofer Institut für Graphische Datenverarbeitung (IGD), 64283 Darmstadt, Germany*

[2]*vsonix GmbH, 64287 Darmstadt, Germany*

Correspondence should be addressed to Gregor Heinrich (gregor.heinrich@igd.fraunhofer.de)

ABSTRACT

We present a platform for real-time transmission of immersive audiovisual impressions using model- and data-based audio wave-field analysis/synthesis and panoramic video capturing/projection. The audio subsystem considered in this paper is based on microphone arrays with different element counts and directivities as well as weakly directional loudspeaker arrays. We report on both linear and circular setups that feed different wave-field synthesis systems. In an attempt to extend this, we present first findings for a data-based approach derived using experimental simulations. This data-based wave-field analysis/synthesis (WFAS) approach uses a combination of cylindrical-harmonic decomposition of cardioid array signals and enforces causal plane wave synthesis by angular windowing and a directional delay term. Specifically, our contributions include (1) a high-resolution telepresence environment that is omnidirectional in both the auditory and visual modality, as well as (2) a study of data-based WFAS realistic microphone directivities as a contribution towards for real-time holophonic reproduction.

1. INTRODUCTION

Telepresence environments have long been investigated in research and real-world applications. By recording a situation in different perceptual modalities in a remote location and reproducing it locally, this technology "enables people to feel as if they are actually present in a different place or time" [10], mostly by transporting important audiovisual cues, which are complemented in many cases by real-time interaction and haptic feedback.

Traditionally, telepresence has been dominated by visual approaches, and the role of the auditory modality was often reduced to enrichment of the video channel with relatively modest quality requirements. This is remarkable because conversely, throughout the history of audio engineering, one of its predominant goals has been (and

still is) "immersion" of the listener into a recorded auditory scene – something that can well be interpreted as auditory telepresence.

To establish a balance between the visual and the auditory modality is the rationale behind the development of our experimental hardware/software telepresence platform, "Omniwall". It integrates high-definition visual telepresence approaches and works towards adequate auditory telepresence.

This article is aimed at introducing the Omniwall approach and to report on our work in progress especially in the audio part. Specifically, Section 2 introduces the scenarios considered and general approaches taken to implement telepresence in them, while the rest of the paper focusses on the audio subsystems developed. Section 3 reviews the model-based audio acquisition and corresponding wave-field synthesis methods implemented in a hardware prototype, including some empirical results. With model-based, we refer to systems that use geometric information and separate source signals to render a wave field according to a wave-propagation model. As a complement to this, in Section 4 we present a design for a data-based wave-field analysis and synthesis system, which we verify with some simulative experiments. Finally, in Section 5, we draw general conclusions and discuss future work.

2. OMNIWALL AUDIOVISUAL TELEPRESENCE

The Omniwall platform can be considered a multi-array approach to audio and video scene recording and reproduction. Functionally, the system consists of an acquisition module, a transmission module and a rendering module, which all handle audio and video information in real time. As deployment environments, we consider two scenarios:

- Panoramic scenario: This corresponds to immersion of the local party into a remote space using an omnidirectional canvas and respective audio reproduction. Such approaches naturally re-enact the surrounding properties of the human perception with a full 360° azimuth of both auditory and visual modalities of the remote scene. [1]
- Planar scenario: This corresponds to typical office conferencing systems where a screen or larger wall provide a "window" to the remote communication party or parties. A special, more telepresence-oriented case of such a setup is a "shared reality" where the remote space is projected locally as an extension of the local space, e.g., with a table that remote and local conferencing partners virtually sit next to each other.

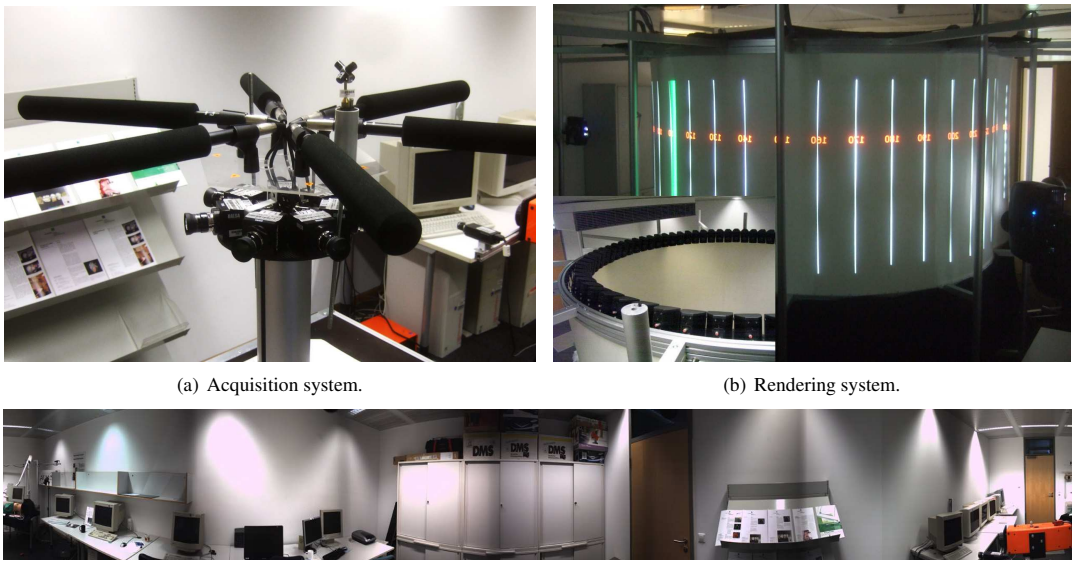
[1] Of course, the visual field of view is localised in azimuth, which is why panoramic visual projection can be interpreted as a means of maximally unobtrusive interaction.

To address these two scenarios and corresponding application domains, we have developed one software system with two different hardware setups: a linear and a circular (or cylindrical) setup. In the following we will briefly describe that environment and the utilised technologies and approaches, whereas the audio wavefield based approach will then follow in detail in sections 3 and 4.

2.1. Circular Omniwall

The circular telepresence system is based on 360° (panoramic) audio-visual acquisition and rendering. There exist many approaches in the literature, covering aspects like tiled projection on different geometries, camera cluster based video recording or the integration with wavefield analysis/synthesis, e.g., [13, 1, 9]. For the panoramic scenario, we chose to use a radial symmetric setup both for the video and the audio part: A camera ring and a corresponding 360 degree projection canvas constitute the visual reproduction system. In the same way the microphones and speakers are aligned (see Fig. 1(a) and 1(b)).

Video branch. The video canvas has a diameter of 2.25 m (7.07 m circumference) and a height (canvas) of 1.40 m. It offers enough space for 5 people standing in its center. Using a circular array of 7 HD cameras and one of 8 HD projectors as well as real-time panoramic stitching techniques, the video branch of this system captures and reproduces a 10 megapixel, 25 fps high-definition omnidirectional video impression of the original space and renders it to a 360° canvas, which surrounds users around ear/eye height. The projection canvas is modelled using a cylinder and the corresponding distortion applied to the video in real-time. In addition, video calibration software exists to align the 8 tiles of the projectors. The perceptual quality of the distortions due to the imperfect canvas surface are, from a perceptual perspective, acceptable for video content with mainly heterogenous visual appearance.



(a) Acquisition system.

(b) Rendering system.

(c) Snapshot of panoramic rendering feed.

Fig. 1: Circular Omniwall telepresence environment.

Audio branch. Using a circular surround microphone array design with 6 directional transducers, an ambiemt impression is captured at the remote location. Within the video canvas, the acoustic field around the remote location is reproduced by way of a circular loudspeaker array of 70 loudspeakers integrated into the canvas (plus 2 subwoofers), using model-based wave-field synthesis (see Sec. 3).

As a complement to the model-based approach, the omnidirectional Omniwall serves as a scenario for data-based wave-field synthesis based on a measurement with a relatively compact circular array (e.g., 50 cm diameter and 48 microphones), which will be described in Sec. 4.

2.2. Linear Omniwall

Although the circular telepresence setup potentially allows a high degree of audiovisual immersion, its space requirements are excessive for many real-world applications. Therefore, a specific version of the Omniwall has been developed that relies on the "shared space" concept outlined above, i.e., the extension of one conferencing room by a projection of the remote space, which can in principle be created symmetrically. The setup of the linear Omniwall is presented in Fig. 2(a) and 2(b). The basic principle we follow for sound recording is based on approaches like the ones presented in [18, 23, 2]. A visual localisation of faces in camera images serves as input for a microphone array that can be steered into the required directions. The seperated voice recordings can then be reproduced at the remote location.

Video branch. The visual channel of the linear Omniwall system consists of a video camera and a corresponding projector setup at the remote location. This connection can easily be replaced with a suitable video encoding stage and network streaming (which is more difficult for the omnidirectional case due to its high bandwidth requirements). The participants of the video conference, which are basically treated as speakers or audio sources, are localised in the camera image using a face detection approach based on [22]. Face locations are associated with audio source locations and can be used to steer the corresponding beamformers into the required direction. In comparison to [2], our system can detect and extract multiple sources at once. Once the individual sources are extracted, the signals are transmitted over the audio connection. The information on where the sources are located is continously updated from the camera images and sent to the remote WFS rendering location over a previously established network connection.

(a) Linear Omniwall, acquisition system. (b) Linear Omniwall, rendering system.

Fig. 2: Linear Omniwall telepresence environment.

Audio branch. The audio part relies on a model-based wave-field synthesis system similar to the circular case described above, now however featuring a linear array of 32 speakers (Canton CD10). In Fig. 2(b), this array is covered with black cloth and can be seen below the video projection.

On the acquisition side, we take information from the visual source localisation (face detection) to steer a linear array of microphones using beamforming methodology. The beamformer array consists of 23 logarithmically spaced cardioid microphones (AudioTechnica Pro45S2) with a total length of 1.8 m. In Fig. 2(a), the array is covered in cloth similar to the speaker array. The beamforming approach is described in section 3.3.

3. MODEL-BASED AUDITORY TELEPRESENCE

Model-based auditory telepresence relies on separate sources and position data to render a remote scene, and this section will describe our approach of combining different means of audio acquisition with model-based wave-field synthesis (WFS) to realise this telepresence method. We will start with a review of the concept of WFS and some specificities of our approach. Subsequently, we will discuss the input data that we use as WFS sources.

3.1. Wave field synthesis

Wave-field synthesis [4] is a technique that allows the reproduction of a wave field in a listening space bounded by loudspeakers with spatio-temporal properties that are, in principle, identical to those in a recording space. In theory, there is no restriction of the listening area to a sweet spot as with stereophonic and surround approaches.

Principle. The physical basis of WFS is the Huygens-Fresnel principle: A wave front from a (primary) source can be completely described by a superposition of waves emitted from (secondary) sources that are located at every point on a wave front of the primary source and oscillate in phase with it. A mathematical description of this is the Kirchhoff-Helmholtz integral, which states that, given a source-free volume, knowledge of pressure and normal particle velocity of the wave field at its boundary surface implies knowledge of the sound pressure at any point in its interior. In other words, in principle it is possible to determine the sound field at the surface of an arbitrary space and reproduce it (in a different time or place) by exciting a congruent surface with exactly the same distributions of pressure and particle velocity.

Viable systems. It is clear that neither the measurement nor the excitation surface can be implemented as a technical system. Therefore, the theoretic approach must be simplified. Literature e.g., [17, 7, 5, 3, 11, 27, 19, 12] suggests several measures to arrive at realisable properties for real-world systems, including:

- Dimensional reduction: The three-dimensional case of an enclosed volume is difficult to build and control in sufficiently large dimensions to place an audience in. Therefore, a restriction to two dimensions

is done, leading to a line as the boundary of the enclosure, which can be placed at the height of ears of the audience. As a result, the distance attenuation of the reproduced wave field is different from that field rendered by a three-dimensional setup and can be adjusted.

- Source pattern restriction: Either monopole (pressure) or dipole (particle velocity) secondary sources can be used to reproduce the field. Usually monopole sound sources are chosen, which simplifies the Kirchhoff-Helmholtz integral to a Rayleigh integral of type II. As a result, the reproduced field outside the enclosure is not reproduced correctly, which is accepted in most cases.

- Spatial discretisation: A continuous excitation of the recording space boundary is not possible due to the minimal geometric dimensions of the loudspeakers. Therefore a discretisation of the boundary excitation is introduced. As a result, wave fields can only be reproduced with wavelengths that are at least double the distance between speakers. Above this, spatial aliasing occurs, i.e., the phase and amplitude distribution of the synthesised field does not correspond to the original field, which results in artefacts. However, humans seem to be not very sensitive to these phenomena.

Further, the way how the sound field at the surface is determined in practice leads to a distinction between *model-based* and *data-based* WFS: Model-based WFS simplifies this measurement by using explicit knowledge on the primary sources with respect to their geometric properties and separate signals, modelling wave propagation through space, while data-based WFS uses field measurements from a small portion of the field, not requiring geometric knowledge or separate source signals but requiring means of wave field extrapolation.

To this article, both approaches are relevant, and we will continue with model-based WFS in this section and present a data-based counterpart in Section 4.

3.2. Model-based WFS realisation

For our telepresence approach, we implemented a model-based WFS system using the simplifications discussed above. Using Verheijen's derivation of model-based WFS via the Rayleigh II integral [21], from each primary source p (signal modelled) to each secondary source q (loudspeaker), a transfer function W_{pq}, also referred to as WFS operator, can be constructed:

$$W_{pq}(\omega) = H(\omega, \lambda_p)\, A_q(\vec{x}_p)\, G_p(\omega, \varphi_{pq}) \exp(-jk(\lambda_p r_{pq} + r_0)) \,, \quad (1)$$

which consists of:

- a filter $H(\omega, \lambda_p)$ that depends on the source region $\lambda_p = \pm 1$ (p inside or outside the array),
- a position-dependent gain factor, $A_q(\vec{x}_p) = f(1/\sqrt{r_{pq}}) \cos\varphi_{pq}$ for $\varphi_{pq} < \pi/2$ and 0 otherwise, where r_{pq} is the distance between primary and secondary sources and φ_{pq} is the angle between their normals at q,
- the directivity of the primary source, $G_p(\omega, \varphi_{pq})$, as well as
- a position-dependent phase shift.

The phase shift term is simply the wave propagation delay, $d(t) = \delta(t - (\lambda_p r_{pq} + r_0)/c)$ where r_0/c is a modelling delay that ensures causal behaviour for "inner sources", i.e., sources projected into the space within the array. The possibility of such inner sources is one of the advantages of WFS because it can be used to let listeners perceive sounds directly in front of them or beside them. Technically, if $\lambda_p = -1$, most importantly the sign of the delay term is reversed, leading to a time-reversed wave that converges in one point in front of the array and subsequently travels as if originating from that point. For the converging part, the additional modelling delay is required.

The structure of the transfer function W_{pq} can be efficiently implemented, as shown in Fig. 3[2], which includes a filter engine that executes all the processing necessary to derive Q loudspeaker signals from P primary source signals. In order to avoid discontinuities in the output signal for moving primary sources, processing is implemented time-variant, naturally leading to a Doppler effect for fast sources as the delay is continuously updated.

This update calculation is done by the parameter engine that processes incoming geometrical data. In addition to Eq. 1, a special case is considered by the module: For primary sources close to the array, the dependence of

[2] A rectangle around a block structure with a number in the lower left corner denotes how many parallel structures exist.

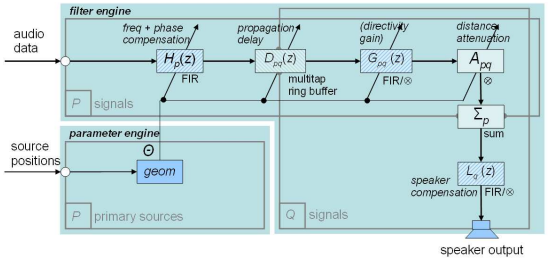


Fig. 3: Model-based WFS system.

$A_q(\vec{x}_p)$ on $1/\sqrt{r_{pq}}$ leads to singularities, and a smoothing heuristic is introduced, effectively implementing vector-based amplitude panning [16] for the limiting case of a primary source on the array radius.

The model-based WFS system operates in both circular and linear telepresence setups where it is fed with different types of source signals. In the following, the different systems that deliver these are described.

3.3. Beamforming-based acquisition

For audio acquisition in the linear Omniwall, two standard approaches [14, 24] for microphone beamforming are employed to spatially separate recording of multiple conference attendees in our scenario, whereas we are able to run the system in three different operation modes:

- Standard Least-Squares: The required spatio-temporal response (target response) of the array is approximated for the given FIR-Filter coefficient using a standard linear Least-Squares optimization.
- Steerable Least-Squares: The filter coefficients are transformed once to a new functional basis (using Legendre polynoms), which allows for simple steering, and transformed back on demand.
- Coefficient look-up: The coefficents are computed offline in advance, for a certain quantization of angles (and distances) and are stored in memory. This method is most efficient, since it involves no computation at all besides multichannel filtering.

The individual source signals are rendered at the remote location in the WFS-based setup using a pre-defined distance. The basic principle of integrating beamforming and WFS is motivated by prior work like presented in [2]. Similar to their system, we utilize visual information (face localisation) to guide audio acquisition.

3.4. Homogeneous surround microphone

Opposed to the linear Omniwall where communication and spatial discrimination are enabling factors, auditory reproduction in the circular Omniwall has no preferred listening direction and should have a reduced dependence on the listening position. In other words, here "true" telepresence criteria apply: the unweighted reproduction of a remote scene in an expanded area in the projection environment. Compared to common recording techniques in audio engineering, this rather corresponds to environmental recordings than multichannel surround recordings (cf. [25, 26]).

As a work in progress, we have constructed an array of 6 supercardioid microphones on an array, as shown in Fig. 1(a). The idea is similar to the IRT microphone cross used for ambience recording [26], with a finer division of the azimuth, though.

On the reproduction side, the signals are fed to 6 virtual primary sources centered in the WFS plane. The radius of these virtual sources can now be used as a parameter to optimise the acoustic impression in the Omniwall, starting at remote sources/plane waves (with appropriate compensation of the distance attenuation), using primary sources on the array to simulate panning/direct rendering, or rendering these channels as inner sources.

3.5. Some empirical results

We are currently in the process of performing empirical tests on both systems and state here preliminary results.

The beamforming approach to acquisition discussed in Sec. 3.3 turned out to be viable, especially for speech, which the beamformer is especially adapted to. We tested the system on a subjective basis by having groups of three test persons sitting on the chairs in the acquisition room. While sitting, the persons engaged in a conversation and moved naturally, except ensuring that their face be visible to the camera. It turned out that under these conditions the face detection algorithm / beamformer system delivered stable audio and positional data to the wave field synthesis.

On the rendering side, five persons were assessing the perceptual quality of the beamforming approach. Speech was very well perceived, and the channel separation between the different beamformers associated with the speakers created a convincing spatial impression matching the video screen (although actual measurements on

the beamformer indicated only between 10 and 15 dB channel separation). We further tested speech with additional non-speech structured noise (drum sounds) as well as concurrent speech. It turned out that the intelligibility of the perceived speech was increased using the directional beamformers, with a wider separation of the WFS sources (artificially moving the primary sources) clearly improving results.

We further varied the depth of the WFS primary sources (that should match the perceived distance to the chairs on the video screen) from far fields behind the array to inner source fields. For the linear array, we were not able, though, to identify an influence of the WFS source depth on speech intelligibility. Therefore, our preliminary result includes that there is no advantage of WFS over panning for the communication scenario.

For the circular Omniwall setup, while generally a spatial impression of the remote location can be well perceived, concrete tests have not been conclusive so far. In the final version of the paper, these tests are being added.

4. DATA-BASED AUDITORY TELEPRESENCE

Opposed to the model-based approach to telepresence, in data-based telepresence no "metadata" on the acoustic sources are available, nor are there separate acoustic source signals. The idea of data-based WFS is rather to analyse an acoustic field in one location regardless of its content and reproduce it in another location as exactly as possible.

Such reproduction in fact was on of the goals of audio reproduction from its beginnings. Mostly, however, reproduction in one reference point, such as stereophony, Ambisonics and HRTF-based methods, was considered a viable alternative to actual "holophony", i.e., reproduction of a complete acoustic field, which was considered intractable.

With recent processing capabilities sufficient to handle even large numbers of microphone and speaker signals, holophonic approaches do not remain unrealistic, and especially in telepresence applications, the idea to have a complete copy of the remote acoustic field is compelling.

In this section, we will attempt exactly to combine results for data-based auditory telepresence by combining methods of wave-field analysis and wave-field synthesis for usage in omnidirectional telepresence systems: "wave-field analysis and synthesis" (WFAS).

Wave-field analysis [5, 12] has been used in the literature mainly for recording of room impulse responses. Here large arrays are constructed or single transducers displaced in multi-measurement setups. In telepresence, usage of such setups would be prohibitive, and more compact designs are preferrable. The circular setup used here should therefore have (1) a small diameter and (2) a moderate number of transducers, two requirements which are antagonists to the requirements of bandwidth and field reproduction quality.

In our ongoing work we have established a simulation environment to predict the behaviour of our circular Omniwall array with different wave-field analysis setups. For our system, we consider all three simplifications of WFS theory described in the Sec. 3.1, and in the next sections we will first review some background and subsequently establish a concrete WFAS operator in analogy to the WFS operator discussed in Sec. 3.1.

4.1. Background

Hulsebos and colleagues have shown [12] that for circular microphone array geometries, fields incident from arbitrary azimuths can be extrapolated in the plane using a cylindrical harmonics decomposition and its relation to the plane-wave decomposition. In their work, they consider paired coincident ideal monopole and dipole elements on the array, with infinitesimal separation. In later work [8], de Vries and colleagues extended this to cardioid microphones, again with ideal assumptions.

Much of the theory can be taken from this work. What is missing for practical WFAS systems is on one hand the consideration of real-world directivities and viable microphone sizes, as (smoothed) measurements strongly deviate from the ideal assumptions. On the other hand, a way to reproduce the field in the listening room without cancelling out waves in a circular enclosure where the problem of causality of the synthesised wave fronts with respect to the original wave occurs. Here, angular windowing as proposed by Spors [20] may filter speaker outputs that emit waves with direction opposite to compared the original wave, i.e., the rear speakers.

4.2. A WFAS operator

To obtain our WFAS operator, we adapt the findings from the mentioned literature to our requirements described above.

Analysis. We consider a circular microphone array of radius R_m with M microphone pairs consisting of

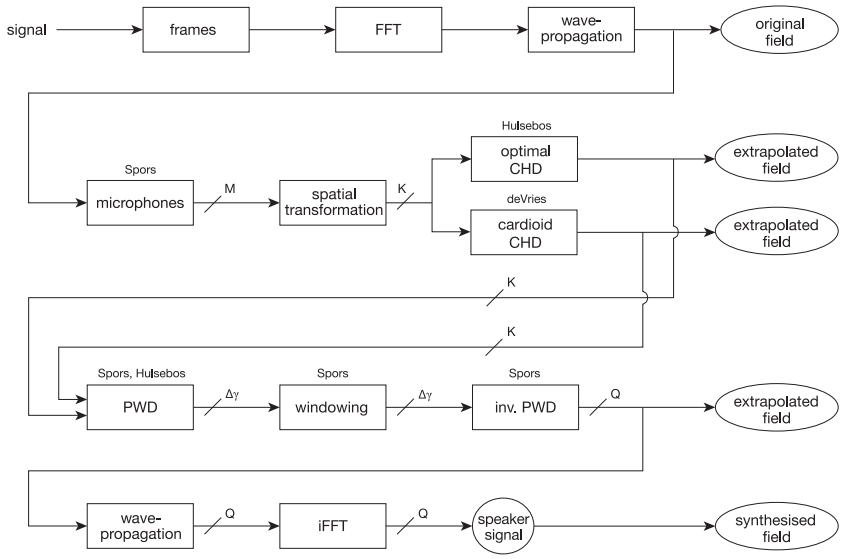


Fig. 4: Signal processing scheme for WFAS approach.

monopole and coincident dipole elements with a radial orientation. The corresponding measurement signals are the pressure $p(m,t)$ and the normal velocity $v(m,t)$ at the m'th pair.

In a first step, the temporal and angular Fourier transforms of the microphone signals are taken:

$$P_{k_\vartheta}(\omega) = \frac{1}{2\pi}\sum_{m=1}^{M} P(m,\omega)\exp(-jmk_\vartheta/2\pi) \tag{2}$$

$$V_{k_\vartheta}(\omega) = \frac{1}{2\pi}\sum_{m=1}^{M} V(m,\omega)\exp(-jmk_\vartheta/2\pi) \tag{3}$$

where k_ϑ can be considered the angular wave number and the constants M and R_m are considered only implicit arguments for notational simplicity. Reconstruction of the original measurements is possible using:

$$P(m,\omega) = \sum_{k_\vartheta=-K}^{K} P_{k_\vartheta}(\omega)\exp(jmk_\vartheta/2\pi) \tag{4}$$

$$V(m,\omega) = \sum_{k_\vartheta=-K}^{K} V_{k_\vartheta}(\omega)\exp(jmk_\vartheta/2\pi)\,, \tag{5}$$

which is exact for $M \to \infty$ and $K \to \infty$. Of course, it is neither possible to obtain an infinite number of angular orders nor to have an infinite number of measurement points. In fact, angular sampling requires that for every order of the angular transform (which corresponds to a periodical function around the array circumference), at least two measurement points are required, leading to the sampling condition of $M > 2K$.

The Fourier-transformed signals can now be developed into a circular harmonics decomposition (CHD) [12]:

$$\begin{aligned} P_{k_\vartheta}(\omega) = {} & \mathcal{M}^{(1)}_{k_\vartheta}(\omega)H^{(1)}_{k_\vartheta}(kR_\mathrm{m}) \\ & + \mathcal{M}^{(2)}_{k_\vartheta}(\omega)H^{(2)}_{k_\vartheta}(kR_\mathrm{m}) \end{aligned} \tag{6}$$

$$\begin{aligned} \varrho_0 c V_{k_\vartheta}(\omega) = {} & \mathcal{M}^{(1)}_{k_\vartheta}(\omega)H^{(1)\prime}_{k_\vartheta}(kR_\mathrm{m}) \\ & + \mathcal{M}^{(2)}_{k_\vartheta}(\omega)H^{(2)\prime}_{k_\vartheta}(kR_\mathrm{m}) \end{aligned} \tag{7}$$

where $\mathcal{M}^{(1,2)}_{k_\vartheta}(\omega)$ are the circular harmonic coefficients for incoming and outgoing waves and $H^{(1,2)}_{k_\vartheta}(kR_\mathrm{m})$ are the Hankel functions of the first and second type and order k_ϑ.

However, practical systems built with such an approach would become overcomplex because a high number of transducer pairs are required, and one of the main questions in our contribution is whether similar results can

be achieved with single signals from realistic microphones replacing the separate signals of the ideal coincident pairs.

Recent work by de Vries and colleagues [8] goes a step in that direction: an approach for an ideal cardioid characteristic is given based (1) on the theoretical view of a cardioid as the average of a monopole and a dipole, as well as (2) the assumption that the halfspace within the array be source-free. These two conditions combined lead to two simplifications: (1) only one signal is necessary per array position, and (2) incoming and outgoing waves can be set equal at the array radius, $\{P, V\}^{(1)}_{k_\vartheta} = \{P, V\}^{(2)}_{k_\vartheta} = \{P, V\}_{k_\vartheta}$, allowing to define an averaged CHD coefficent. Some restructuring of the above equations leads to:[3]

$$\mathcal{M}_{k_\vartheta}(\omega) = \frac{\mathcal{M}^{(1)}_{k_\vartheta}(\omega) + \mathcal{M}^{(2)}_{k_\vartheta}(\omega)}{2} \tag{8}$$

$$S_{k_\vartheta}(\omega) = \frac{P_{k_\vartheta}(\omega) + j\varrho_0 c V_{k_\vartheta}(\omega)}{2} \tag{9}$$

$$= 2\mathcal{M}_{k_\vartheta}(\omega)\left(H^{(1)}_{k_\vartheta}(kR_\mathrm{m})H^{(2)}_{k_\vartheta}(kR_\mathrm{m}) - j(H^{(1)'}_{k_\vartheta}(kR_\mathrm{m})H^{(2)'}_{k_\vartheta}(kR_\mathrm{m})\right) \tag{10}$$

$$= 4\mathcal{M}_{k_\vartheta}(\omega)(J_{k_\vartheta}(kR_\mathrm{m}) - jJ'_{k_\vartheta}(kR_\mathrm{m})) \tag{11}$$

where $S_{k_\vartheta}(\omega)$ represents the angular-temporal Fourier transform of the cardioid microphone signals analougous to $P_{k_\vartheta}(\omega)$.

The CHD representation is compelling because it has a close connection to the plane wave decomposition of the field. This can be used to extrapolate field values from the array positions, which is one of the keys to synthesis of the measured field.

Synthesis. For synthesis, we consider a circular loudspeaker array with radius R_q and Q loudspeakers with an approximate monopole directivity. The synthesis task consists of driving these speakers with the signals of the CHD derived in the wave-field analysis. The main idea here is to extrapolate the CHD representation (which is bound to the microphone radius R_m) to the speaker positions at radius R_q and subsequently ensure that the synthesised field is causal with respect to the original field. That is: no synthesised wave should travel in opposite direction to its corresponding original wave (excluding the case of time-mirrored inner sources for now).

[3] The last simplification is due to Lars Hörchens.

A common way to extrapolate signals from the CHD representation is to use the plane-wave decomposition (PWD). The PWD, $\bar{P}(\vartheta, \omega)$, represents a wave field by a superposition of plane waves from all directions incident at one reference point, theoretically allowing to reconstruct the pressure at any field point:

$$P(r, \vartheta, \omega) = \frac{1}{2\pi}\int_0^{2\pi} \bar{P}(\vartheta', \omega)\exp(-jkr\cos(\vartheta - \vartheta'))\,\mathrm{d}\vartheta' \,. \tag{12}$$

As the analysed field is expressed in terms of circular harmonics, the plane wave decomposition can be easily obtained using the relation [12, 8]:

$$\bar{P}(\omega, \vartheta) = \frac{1}{\pi}\sum_{k_\vartheta=-K}^{K}(-j)^{k_\vartheta}\mathcal{M}_{k_\vartheta}(\omega)\exp(jk_\vartheta\vartheta) \tag{13}$$

where $K = (M-1)/2$.

The final step is now to obtain loudspeaker signals that adhere to the causality condition described above. Following a similar path as Spors [20], we apply an angular window to the partial plane waves, that is, the effect of every plane wave is weighted with its positive angle to the loudspeaker normal in question and set to zero for negative angles, which would correspond to non-causal propagation.

However, to explicitly derive loudspeaker signals, the solution given in [20] appears insufficient if no source position is known. A solution to this is to ensure that the loudspeakers radiate in phase with respect to any partial plane wave, which is achieved by a delay term. Consider an azimuth angle on the array, α, where a loudspeaker is located. A plane wave with the same angle of incidence as α will be excited without delay. The signal belonging to the same plane wave for a speaker located at another azimuth will have to be delayed by the time that the plane wave needs to travel until it reaches the second speaker. This can be expressed by the azimuth difference between the speakers, denoted as γ, which is simply $R_\mathrm{q}/c \cdot (1 - \cos\gamma)$.

To derive the signal for a specific speaker, the contributions of all PWD angles have to be summed up. Combining this with generic results in [20], the pressure distribution on the synthesis array in polar coordinates (R_q, α) is:

$$P(\omega, R_\mathrm{q}, \alpha) = -\frac{jk^2}{4\pi^2}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\bar{P}(\omega, \alpha+\gamma)\cos(\gamma) \cdot \exp(-jkR_\mathrm{q}(1-\cos\gamma))\,\mathrm{d}\gamma \,. \tag{14}$$

In addition to introducing the delay explained above, this integral performs angular windowing using the cosine term $\cos\gamma$ in connection with the integration limits, which ensures that the causality condition is fulfilled. For any loudspeaker on the array q, the driving signal can be finally derived using discretisation:

$$P(q,\omega) = -\frac{jk^2}{4\pi^2}\sum_{\nu=-N}^{N}\bar{P}(\omega,\alpha_q+\nu\Delta\gamma)\cos(\nu\Delta\gamma) \cdot \exp(-jkR_\mathrm{q}(1-\cos(\nu\Delta\gamma))) \quad (15)$$

where $\alpha_q = 2\pi q/Q$ is the actual azimuth angle of speaker q and the stepping $\Delta\gamma = \pi/(2N)$ is chosen according to the resolution of the PWD.

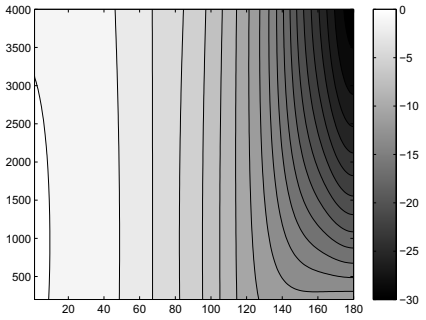


Fig. 5: Interpolated microphone directivity (angle [°] vs. frequency [Hz].

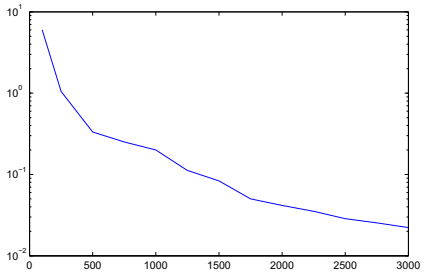


Fig. 6: Measured gain factors.

4.3. Experiments

The goal of this contribution is to assess the quality of field reproduction using the described WFAS approach. One of the central parts of such an assessment is the introduction of realistic side conditions. This includes to take into account non-ideal microphones based on measured directivities.

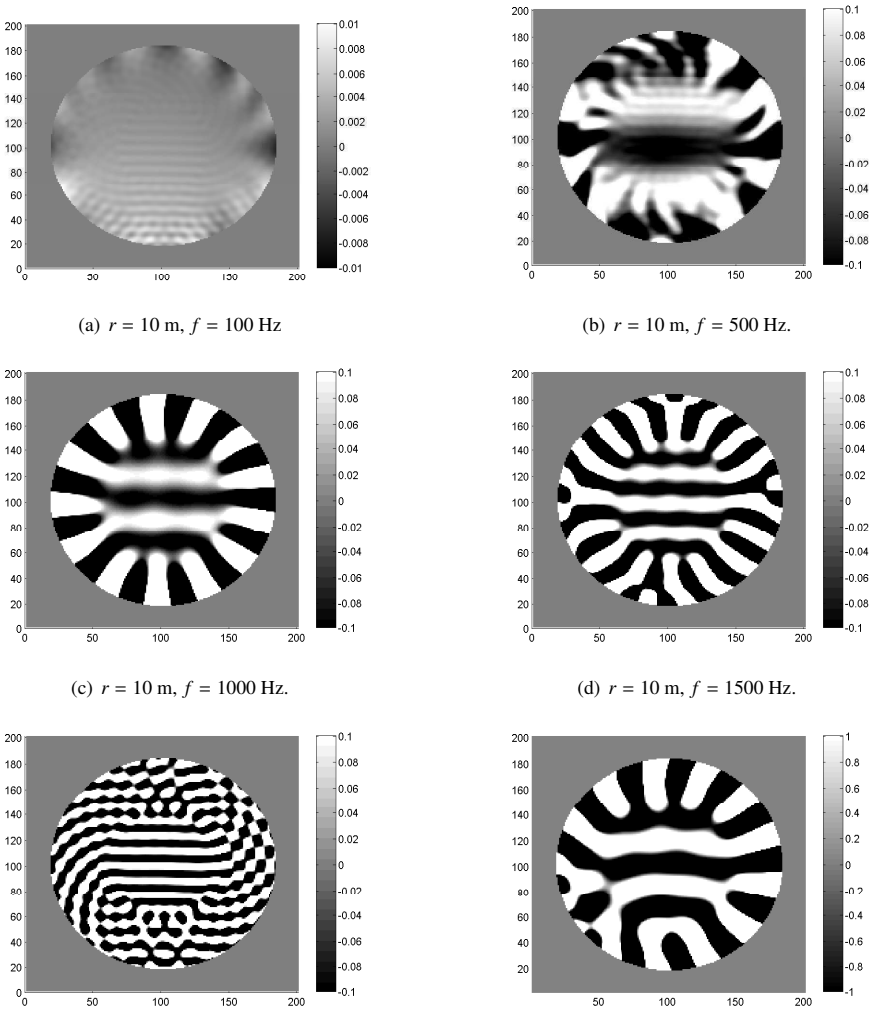
However, instead of taking into account all of the transducer measurements, which leads to a complex model, we restrict our current simulative study to a smoothed model of a real cardioid microphone that captures the deviations from ideal frequency behaviour but is simple to calculate. In Fig. 5, the smoothed directivity pattern of an AudioTechnica Pro45S2, the microphone array used in the linear beamformer setup (see Sec. 3), is presented as a function of angle and frequency, which has been obtained by spline interpolation of the measured directivities.

With this microphone, we have performed a number of simulations. The simulation setup is as shown in Fig. 4, i.e., the wave from a primary point source is propagated to the microphone array where it is processed according to the steps described in the last section. Finally, the secondary field generated by the loudspeaker array as pressure excitation is analysed for its properties. For our simulations, we use a realistic microphone array of with $R_\mathrm{m} = 0.25$m and $M = 47$, which implies an CHD order of $K = 23$. On the rendering side, our existing loudspeaker array is simulated with $R_\mathrm{q} = 1.125$m and $Q = 70$. The main criterion of quality of this secondary field is the difference to the original field. Further, we use the microphone directivity given in Fig. 5.

We first empirically analysed the gain factors necessary to reproduce different frequencies by energy comparison over the center part of the field that the reproduced field has to be scaled with. For the far source, these factors are presented in Fig. 6. Here the strong high-pass behaviour of the system becomes clear, which is similar to that of higher-order Ambisonics (HOA) systems [6], but appears less steep and more irregular, the latter likely due to the non-ideal properties of the microphone directivity.

After normalising synthesis amplitudes, we analysed both far field sources with center distance of $r = 10$m and near field sources $r = 2.5$m, and Fig. 7 shows the resulting instantaneous field pressure for the listening space in a 2.25m array of 70 loudspeakers for different frequencies. The fields are calculated only for the center part of the field, but the general tendency of the reproduced field is well visible. Further, Fig. 8 shows the instantaneous error fields, which are normalised to the RMS pressure at the center.

What can be seen is that the field is correctly reproduced



(a) r = 10 m, f = 100 Hz

(b) r = 10 m, f = 500 Hz.

(c) r = 10 m, f = 1000 Hz.

(d) r = 10 m, f = 1500 Hz.

(e) r = 10 m, f = 3000 Hz.

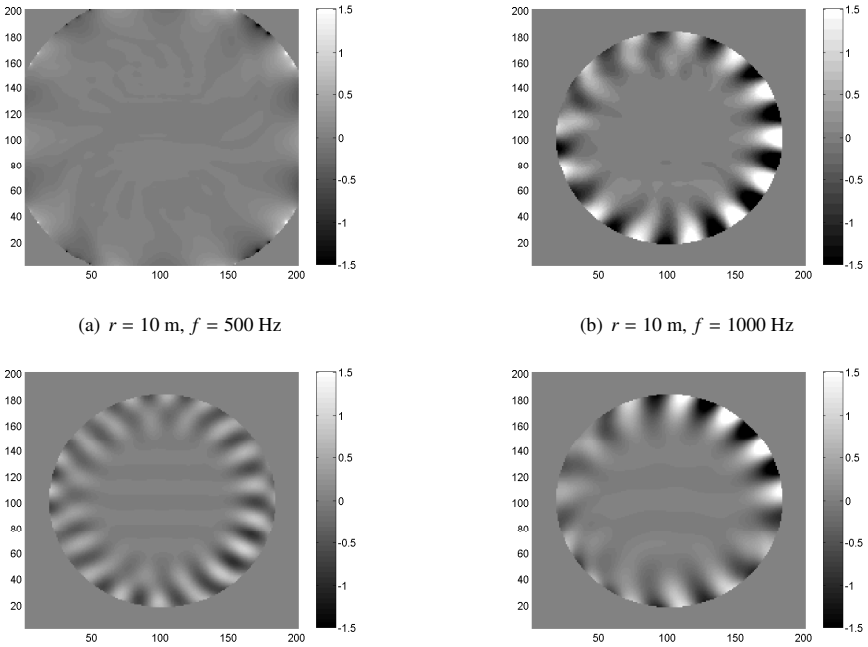
(f) r = 2.5 m, f = 1000 Hz.

Fig. 7: Instantaneous rendered field pressure at different frequencies (x, y-coordinates = cm).

in the center of the array, with normalised error below 10% over an area with approximately the double radius of the microphone array for frequencies around 1kHz. Again, similar to HOA, the region of good reproduction increases with the wavelength (cf. 100 Hz in Fig. 7(a) and 3000 Hz in Fig. 7(e)), and again, from first qualitative evaluation, the dependence on the frequency seems smaller. This is true for both the far and near source.

5. CONCLUSIONS

In this paper, we have presented (1) a reproduction environment for audiovisual scenes in high resolution. For this, we have shown the general setup in two scenarios. Further, we have (2) presented a study of data-based wave-field synthesis method to analyse its viability in the telepresence setup. Opposed to the literature on data-based wave-field synthesis, where the method was



(a) r = 10 m, f = 500 Hz

(b) r = 10 m, f = 1000 Hz

(c) r = 10 m, f = 1500 Hz

(d) r = 2.5 m, f = 1000 Hz

Fig. 8: Normalised instantaneos error field (x, y-coordinates = cm).

mostly used to determine spatial room impulse responses or to determine theoretical results, we explicitly target at a system that is realistically implementable for our telepresence approach, which implies the usage of compact microphone arrays with realistic microphone directivities as well as a general capability to render a field in real time.

Our work, however, is in progress, and we consider most of the reported results still inconclusive. In a later version of this paper, we will update and elaborate our findings. This goes into three directions: First, we are currently experimentally studying the model-based wave-field synthesis in connection with the surround microphone approach, possibly exchanging the microphones by less frequency dependent directivities. As another alternative to this array, we recently have added a first-order Ambisonics system as a comparison for the surround/WFS approach, which requires more in-depth analysis as to source localisation.

Second, concerning the audio acquisition part of the linear Omniwall, we plan to further improve detection of speaker sources by the implementation of a body/people localisation approach, as the continous and stable estimation of source locations is crucial for usability in a real scenario. We also expect to enhance quality of spatially extracted audio by utilizing adaptive beamforming and post-filtering.

Third, regarding the wave-field analysis and synthesis (WFAS) approach proposed in the second part of this paper, the planned study of its properties will include variations of array size and transducer count as well as a more thorough investigation of near field effects (always under the assumption of realistic directivities), with possible compensation similar to that of higher-order Ambisonics (HOA) [6]. In fact, many properties of the WFAS approach seem similar to HOA (likely due to cylindrical harmonics at the core of both methods), and a deeper investigation of the theoretical similarities will be done, possibly in addition considering WFAS a special case of a general theory of sound reproduction (e.g., [15]).

ACKNOWLEDGEMENTS
Parts of this work were supported by the European Union through the Integrated Project "hArtes - a holistic approach to real-time embedded systems".

6. REFERENCES

[1] Jonathan Baldwin, Anup Basu, and Hong Zhang. Panoramic video with predictive windows for telepresence applications. In *ICRA*, pages 1922–1927, 1999.

[2] J. A. Beracoechea, S. Torres-Guijarro, L. García, and F. J. Casajús-Quirós. On building immersive audio applications using robust adaptive beamforming and joint audio-video source localization. *EURASIP J. Appl. Signal Process.*, 2006(1):196, January uary.

[3] A. J. Berkhout, D. de Vries, J. Baan, and B. W. van den Oetelaar. A wave field extrapolation approach to acoustical modeling in enclosed spaces. *The Journal of the Acoustical Society of America*, 105(3):1725–1733, 1999.

[4] A.J. Berkhout. *Applied Seismic Wave Theory*. Elsevier, 1987.

[5] A.J. Berkhout, D. de Vries, and J.J. Sonke. Array technology for acoustic wave field analysis in enclosures. *J.Acoust.Soc.Am.*, 102:2757–2770, 1996.

[6] Jerome Daniel, Rozenn Nicol, and Sebastien Moreau. Further investigations of higher-order ambisonics and wavefield synthesis for holophonic sound imaging. In *Preprints of the 114th AES Convention*, 2003.

[7] Diemer de Vries and Marinus M. Boone. Wave field synthesis and analysis using array technology. In *Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York*, Oct 1999.

[8] Diemer de Vries, Lars Hörchens, and Peter Grond. Extraction of 3D information from circular array measurements for auralization with wave field synthesis. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007.

[9] Kiran J. Fernandes, Vinesh Raja, and Julian Eyre. Cybersphere: the fully immersive spherical projection system. *Commun. ACM*, 46(9):141–146, 2003.

[10] Scott Fisher and Brenda Laurel. Be there here. *InterCommunication*, 1992.

[11] R. Hldrich and A. Sontacchi. Wellenfeldsynthese erweiterungen und alternativen. In *DAGA 2005, 31. Deutsche Jahrestagung fr Akustik, Mnchen*, 2005.

[12] Edo Hulsebos, Diemer de Vries, and Emmanuelle Bourdillat. Improved microphone array configurations for auralization of sound fields by wave field synthesis. *J. Audio Eng. Soc.*, 50(10):779–790, 2002.

[13] S. Ikeda, T. Sato, and N. Yokoya. High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system. *Multisensor Fusion and Integration for Intelligent Systems, MFI2003. Proceedings of IEEE International Conference on*, pages 155–160, July-1 Aug. 2003.

[14] Lucas C. Parra. Steerable frequency-invariant beamforming for arbitrary arrays. *Journal of the Acoustical Society of America*, 119 (6):3839–3847, 2006.

[15] Mark Poletti. A unified theory of horizontal holographic sound systems. *J. Audio Eng. Soc.*, 48(12), 12December 2000.

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

[17] R. Rabenstein and S. Spors. Spatial sound reproduction with wave field synthesis. In *Atti del Congresso 2005, Audio Engineering Society, Italian Section, Como, Italia*, Nov 2005.

[18] K. Shinoda, H. Mizoguchi, S. Kagami, and K. Nagashima. Visually steerable sound beam forming method possible to track target person by real-time visual face tracking and speaker array. *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, 3:2199–2204 vol.3, Oct. 2003.

[19] S. Spors, H. Buchner, and R. Rabenstein. Eigenspace adaptive filtering for efficient pre-equalization of acoustic MIMO systems. In *Proc. European Signal Processing Conference (EUSIPCO), Florence, Italy*, Sep 2006.

[20] Sascha Spors. *Active Listening Room Compensation for Spatial Sound Reproduction Systems*. PhD thesis, University of Erlangen-Nuremberg, 2006.

[21] E. Verheijen. *Sound Reproduction by Wave Field Systnesis*. PhD thesis, TU Delft, 1998.

[22] Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[23] Ce Wang, Scott Griebel, Michael Brandstein, and Bo-June (Paul) Hsu. Real-time automated video and audio capture with multiple cameras and microphones. *J. VLSI Signal Process. Syst.*, 29(1-2):81–99, 2001.

[24] Darren Brett Ward. *Theory and application of broadband frequency invariant beamforming*. PhD thesis, January 01 1996.

[25] M. Williams and G.Le Dû. The quick reference guide to multichannel microphone arrays part 1: using cardioid microphones. In *110th AES Convention in Amsterdam preprint 5336*, 2001.

[26] Joerg Wuttke. Surround recording of music: Problems and solutions. In *Proc. 119 AES Convention*, 2005.

[27] Sylvain Yon, Mickael Tanter, and Mathias Fink. Sound focusing in rooms: The time-reversal approach. *JASA*, 113:1533–1543, 2003.