AN INTEGRATED REAL-TIME SYSTEM FOR IMMERSIVE AUDIO APPLICATIONS

*H. Teutsch, S. Spors, W. Herbordt, W. Kellermann, R. Rabenstein**

University of Erlangen-Nuremberg
Telecommunications Laboratory
Cauerstr. 7, 91054 Erlangen, Germany
{teutsch,spors,herbordt,wk,rabe}@LNT.de

ABSTRACT

In this paper a real-time system for immersive audio applications is presented. Sound sources are recorded using a microphone array whose beam is steered according to the output of an acoustic source localization and tracking system. The output of the beamformer (BF) along with the source position updates are continuously transmitted to a wave field synthesis (WFS) system. By using WFS the sound sources in the recording room are rendered in the reproduction room with the correct spatial cues.

1. INTRODUCTION

Wave field synthesis has received much attention since its introduction in the early 1990s [1] as a technique for exactly reproducing a sound field within a listening area that is dictated only by the loudspeaker setup. Compared to traditional multi-channel sound reproduction techniques such as stereo or '5.1' the listener now seems to be truly immersed by the rendered sound field. Another advantage of WFS is its virtually unlimited flexibility and scalability in terms of loudspeaker setup and auralization possibilities. Note that stereophony and other forms of multi-channel reproduction can be regarded as a subset of WFS.

Systems utilizing WFS for immersive audio applications have been described previously [2] along with rather impressive real-time demonstrations of its capabilities at various exhibitions and conventions. However, none of these approaches has considered live sound recording as an input for the rendering system. The previously presented systems mostly aimed at applications such as cinemas and non-live broadcast scenarios. This contribution tries to fill this void by now allowing applications such as live broadcast scenarios to be implemented utilizing WFS techniques. In this work, sound sources are recorded using a microphone array whose beam is steered following the output of an acoustic source localization and tracking system. The output of the BF along with the source position updates are continuously transmitted to an WFS rendering system.

This paper is organized as follows. Section 2 discusses the microphone array used for sound recording in this work while Section 3 outlines the idea of acoustic source tracking and briefly describes the tracking system used in our setup. Section 4 gives a short overview of WFS as a technique for rendering of 3D sound fields. Finally, Section 5 describes our integrated real-time system followed by some conclusions and outlines for future work in Section 6.

* This work was partly supported by grants from the European Commission as sponsor of the CARROUSO project, and from Intel Corp.

2. MICROPHONE ARRAY FOR ACOUSTIC SOURCE RECORDING

Microphone array technology has been traditionally utilized for speech pickup applications such as teleconferencing and seamless hands-free human-machine interfaces (see, e.g., [3] and references therein). The frequency-band of interest for these types of applications is often chosen as the one used for conventional telephony, i.e., 300-3400 Hz. Several adaptive and non-adaptive microphone array designs have been proposed that operate quite well in this frequency range [3]. While speech-based applications pose a challenging problem by themselves, this work tries to go one step further by considering the application of microphone array technology to high-quality sound recording. Among speech, musical instruments are the targeted sound sources to be recorded and spatially separated. Microphone arrays have the potential advantage of relieving the actors/musicians of having to carry or wear close-up microphones. Thus they allow for a so-called 'hands-free' recording environment.

By adhering to the constraints of very-large bandwidth and distortionless real-time operability, we have kept our focus on 'signal-independent beamforming' arrays in this work. The following section briefly describes the array design technique used.

2.1. Array Design

Consider a plane wave emitted by a source n associated with the directional wavenumber $\mathbf{k}_n \in \mathbb{R}^3$. The output of a one-, two- or three-dimensional BF b_n with M microphones fixed at positions $\mathbf{p}_m \in \mathbb{R}^3$ can then be written as

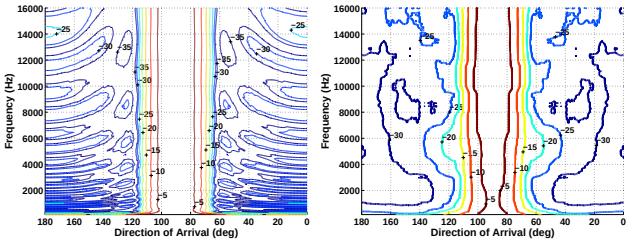
$$b_n = \sum_{m=0}^{M-1} w_m \, e^{j(\mathbf{k}-\mathbf{k}_n)\cdot\mathbf{p}_m}, \qquad (1)$$

where w_m is the FIR filter attached to microphone m and '$\cdot$' denotes a scalar product. Note that the dependence of b_n on frequency and direction of sound incidence, and the dependency of w_m on frequency has been dropped for notational convenience. Note also that N sources can be combined into a BF vector $\mathbf{b} = [b_1, b_2, \ldots, b_N]$.

As can be deduced from Eq. (1) there are two parameters, w_m and $\mathbf{p}_m$, that can be utilized for the BF design. Keeping the high-quality paradigm in mind there are several constraints to be met for the array design including a constant beamwidth of the array over the entire frequency range as well as a favorable sidelobe structure. Following the discussions presented by Doles [4], Ward [3], and van der Wal [5], in order to fulfill the constraint of

constant beamwidth one has to make sure that the array's aperture scales with frequency. This means that for the BF design w_m and $\mathbf{p}_m$ have to be optimized simultaneously and that constant beamwidth cannot be obtained by an array structure that utilizes linearly spaced microphones only. See [3], [4], and [5] for a detailed treatment of this problem.

For the application to wave field synthesis the array should be able to provide spatial separation down to about 150 Hz. Thus, a one-dimensional 26-element logarithmically-spaced microphone array of 2.6 m length has been designed by combining ideas presented in the aforementioned citations with a directivity pattern as shown in Fig. 1.



(a) *simulated* (b) *measured*

Figure 1: *Directivity of the 26-element microphone array*

As can be seen, the measured performance of the array closely follows the simulated one. The spatial selectivity seems to be attractive for high-quality sound acquisition. A section of the microphone array setup as used in the integrated real-time system (see Section 5) is shown in Fig. 2.

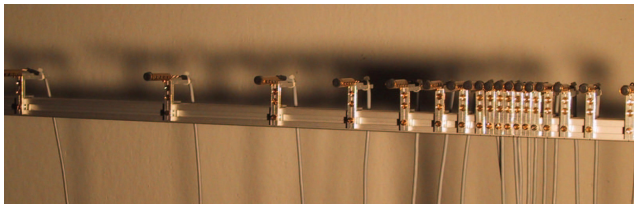


Figure 2: *Photograph of the microphone array system*

3. ACOUSTIC SOURCE LOCALIZATION AND TRACKING

Our scenario should enable the users to move freely within the recording room. Sound recording with microphone arrays, which exploit spatial information, requires localization and tracking of the sound sources in order to steer the array to the correct position(s). In our setup, the 3D source positions are estimated acoustically using a microphone array consisting of four transducers [3] which are somewhat heuristically arranged in two dimensions.

Acoustic source localization particularly has to deal with the inherent challenge of reverberation in typical acoustic environments. Strong reflection paths are spuriously interpreted as the direct signal path leading to erroneous position estimates.

In our setup, we address these problems by a four-step procedure as shown in Fig. 3. First, the microphone signals are fed into a signal activity detector which distinguishes between activity of the desired sound source and the presence of background noise only by adaptive thresholding. Here, we assume that the background noise possesses a slowly time-varying energy relative to the one of the desired sound source. Only if the short-time energy of a frame of microphone signals is above a given threshold, the frame is used for localization. The threshold is adjusted according to a long-time average of the sensor data's energy.

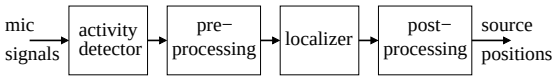


Figure 3: *Flow-graph of the source localization system*

Then, the sensor signals are pre-processed by a cepstral pre-filter which removes dominant indirect propagation paths [6]. Phase-sensitive noise suppression follows as a second pre-processing step.

The pre-processed sensor signals are then presented to the localizer. In general, acoustic source localization is based on the estimation of relative time delays between the source position and the sensor positions and on a geometrical interpretation of the estimated time delays. For the estimation of the propagation delays, we implemented various algorithms which have been previously presented in the literature dealing with robust time-delay estimation in reverberant environments. For the highest possible flexibility, our realization utilizes either adaptive eigenvalue decomposition (AED) [7], generalized cross-correlation (GCC) using various normalization terms, or steered response power (SRP) methods [3]. AED is based on the blind estimation of room impulse responses between the source and the sensors, GCC uses (normalized) cross-correlations between the sensor signals, and SRP measures the energy of BF outputs where the BF is steered to possible source positions. For a comparison of the source localization algorithms the interested reader is referred to the literature, e.g., [3].

Finally, the post-processing unit is aimed at removing impossible source position estimates, i.e., severe outliers in the estimation process. The final position estimates are then smoothed by a simple median filter which also reduces the impact of moderate outliers.

Note that the time-delay estimation techniques summarized here are not able to deliver reliable estimates for two or more simultaneously active sound sources. Therefore, only one active sound source in the recording room is considered in this work. However, the authors believe that an incorporation of voting strategies, such as the one presented in [8], will help resolving the ambiguity problem.

4. WAVE FIELD SYNTHESIS FOR RENDERING OF ACOUSTIC SOURCES

The theory of WFS has been initially developed at the Technical University of Delft over the past decade [1]. This section gives a short overview of the theory as well as on rendering methods and wave field analysis.

4.1. Theoretical Background

WFS is based on the Huygens' principle. Huygens stated that any point of a wave front of a propagating wave at any instant conforms to the envelope of spherical waves emanating from every point on the wavefront at the prior instant. This principle can be used to synthesize acoustic wavefronts of an arbitrary shape. Of course, it

is not very practical to position the acoustic sources on the wavefronts for synthesis. By placing the loudspeakers on an arbitrary fixed curve and by weighting and delaying the driving signals, an acoustic wavefront can be synthesized with a loudspeaker array. Figure 4 illustrates this principle.

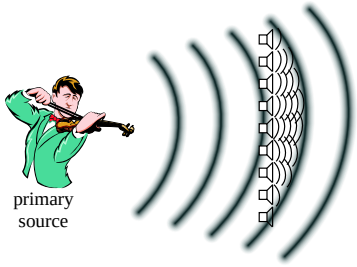


Figure 4: *Basic principle of wave field synthesis*

The mathematical foundation of this rather illustrative description of WFS is given by the Kirchhoff-Helmholtz integral, which can be derived by using the wave equation and the Green's integral theorem [9]. The Kirchhoff-Helmholtz integral states that for any listening point within a source-free volume V the sound pressure can be calculated if both the sound pressure and its gradient are known on the surface S enclosing the volume. This principle can be used to synthesize a wave field within a volume V by applying the appropriate pressure distribution and its gradient on the surface. However, two essential simplifications are necessary to arrive at a realizable system: Degeneration of the surface S to a line and spatial discretization. Performing these steps the so-called Rayleigh integrals can be derived [1]. The Rayleigh I integral states that a pressure field may be synthesized by means of a monopole distribution on a line. Using this result a WFS system can be realized by mounting electro-dynamic loudspeakers in a linear setup surrounding the listening area. Figure 5 shows the

Figure 5: Photograph of the WFS system

24-element loudspeaker array setup utilized in this work. Up to now we assumed that no acoustic sources lie inside the volume V. The theory outlined above can also be extended to the case where sources lie inside the volume V [1].

4.2. Rendering Techniques

In general, the vector of L loudspeaker input signals $\mathbf{q}[k]$ can be expressed as a convolution of a measured or synthesized filter matrix $\mathbf{W}[k]$ with the vector of N virtual source signals $\mathbf{s}[k]$:

$$\mathbf{q}[k] = \mathbf{W}[k] * \mathbf{s}[k], \qquad (2)$$

where k denotes the discrete time index, and '$*$' the convolution operator. There are two different approaches to compute the so-called WFS operator $\mathbf{W}[k]$ often referred to as rendering techniques:

1. *Model-based rendering*
   Point sources and plane waves are the most common models used here. These models for the spatial source characteristics are used to calculate the filter matrix $\mathbf{W}[k]$ which consists of simple weights and delays in this case.

2. *Data-based rendering*
   Here, the WFS operator $\mathbf{W}[k]$ for auralization cannot be obtained by simply measuring or simulating the impulse responses from a source to a listener position. The wave field has to be captured in a way such that it yields information on the traveling direction of the sound waves. The filter matrix $\mathbf{W}[k]$ can be calculated offline after the wave field has been recorded with a dedicated microphone array. Post-processing, utilizing wave field analysis techniques, unveils then the desired wave field information [10].

5. SYSTEM INTEGRATION AND REAL-TIME IMPLEMENTATION

This section describes the integration and real-time implementation of the sub-systems BF, acoustic source tracking, and WFS to yield a framework suitable for the realization of a complete transmission system from recording to rendering in immersive audio applications.

Figure 6 shows an example scenario, e.g. a broadcast application, where a single moving primary sound source, as indicated by the arrows, is captured in the recording room. The output of the BF-signal as well as the instantaneous source position are then presented continuously to the WFS system at the reproduction site for spatially correct rendering using techniques as described in Section 4.2. Since the BF signal should, ideally, deliver dry audio signals, a room different from the actual recording room can be auralized by using appropriate sets of pre-recorded room parameters (RP) for obtaining the WFS operator $\mathbf{W}[k]$.

In order for the BF to be able to steer its beam according to the momentary position of the sound source, a communication to the source tracking system has to be established. As indicated by the dashed box in Fig. 6 the single source tracking system can be extended by using more tracking systems whose estimates can be combined by a data fusion center. By using a decentralized Kalman Filter (KF) position estimates obtained by different tracking modalities, for example by acoustical, optical, or infrared means, can be combined into one global source position estimate (see Ch. 10 from [3]). Even if only a single source tracking system is used, as in the present work, using an KF can be advantageous since motion models can be incorporated that help increasing the reliability of the tracking results.

The communication between the individual components works as follows. Following registration of all components with the fusion center, which works as a server, the position updates delivered by the acoustic source tracking system are sent to the fusion center based upon a textual description of the data encoded in KQML (knowledge query and manipulation language) via TCP/IP (transmission control protocol/internet protocol). The smoothed position data are then sent to the BF, using the same communication model, which then steers its beam accordingly. The thus

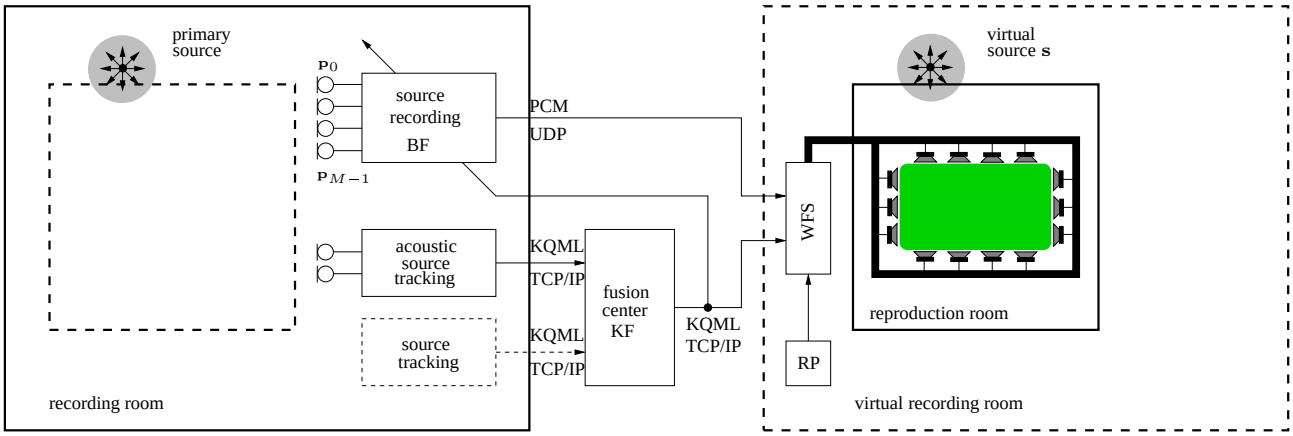


Figure 6: *System integration (only one source shown)*

acquired audio data are presented to the WFS system in standard PCM (pulse coded modulation) format using UDP (user datagram protocol). Audio is recorded using 48 kHz sampling rate and 16 bit sample resolution. The audio data along with the estimated source position are then used by the WFS system to auralize the source within the listening area which is indicated by the shaded area in Fig. 6.

Although not explicitly shown in Fig. 6, the image of the sound source in the recording room could also be captured by a video camera and transmitted to the reproduction room.

Note that the entire system is highly scalable. The number of audio sources recorded, transmitted and rendered solely depends on the available computational power and network bandwidth. At this point the limitation is introduced by the acoustic tracking system that imposes the constraint that only one audio source may be active at a given time.

The entire real-time system runs on four standard PCs, one for each sub-system. All sub-systems have been implemented under the LINUX operating system using the JACK [11] audio interface that offers amazingly high flexibility, high-performance, and low-latency operation. The latency of the entire system is in the range of a few hundred milliseconds. Subjective evaluation has confirmed that a moving sound source captured and tracked in the recording room is rendered in the reproduction room with the correct spatial cues in high sonic quality.

6. CONCLUSIONS

We have presented a real-time framework that integrates source recording using BF, acoustic source tracking, and rendering using WFS. This system could be utilized, e.g., in broadcast applications and tele-teaching scenarios. Future work will include an investigation of real-time capable multi-source tracking algorithms and the integration of other source tracking modalities such as video. Furthermore, efforts will be made to also enable full-duplex operation. In this case, multi-channel acoustic echo cancellation (AEC) will become necessary. It has been recently shown [12] that a combination of AEC and WFS is possible and, given the available computational power, feasible.

7. REFERENCES

[1] Berkhout, A.J., de Vries, D., and Vogel, P., "Acoustic control by wave field synthesis", J. Acoust. Soc. Am., vol. 93, 1993, pp 2764–2778.

[2] Sporer, T., Plogsties, J., and Brix, S., "CARROUSO – An European Approach to 3D-Audio", AES 110th convention, Amsterdam, The Netherlands, May 2001.

[3] Brandstein, M.S., and Ward, D.B., eds, Microphone Arrays: Signal Processing Techniques and Applications, Springer Verlag, 2001.

[4] Doles, J.H., and Benedict, F.D., "Broad-Band Array Design Using the Asymptotic Theory of Unequally Spaced Arrays", IEEE Trans. Ant. Prop., vol. 36, no. 1, Jan. 1988, pp. 27-33.

[5] van der Wal, M, Start, E.W., and de Vries, D, "Design of Logarithmically Spaced Constant-Directivity Transducer Arrays", J. Audio Eng. Soc., vol 44, June 1996, pp. 497-507.

[6] Stéphenne, A., and Champagne, B., "A new cepstral prefiltering technique for estimating time delay under reverberant conditions", Signal Processing, vol. 59, no 3, 1997, pp. 253-266.

[7] Benesty, J., "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization", J. Acoust. Soc. Am., vol. 107, Jan. 2000, pp. 384-391.

[8] Kellermann, W., "A self-steering digital microphone array", Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 5, May 1991, pp. 3581-3584.

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

[10] Hulsebos, E., de Vries, D., and Bourdillat, E., "Improved microphone array configurations auralization of sound fields by Wave Field Synthesis", 110th AES Convention, Amsterdam, The Netherlands, May 2001.

[11] Davis, P., et al., Jack Audio Connection Kit, http://jackit.sourceforge.net.

[12] Buchner, H., Spors, S., Kellermann, W., and Rabenstein, R., "Full-duplex communication systems with loudspeaker arrays and microphone arrays", IEEE Proc. Int. Conf. on Multimedia and Expo, Lausanne, Switzerland, August 2002.