AMBISONICS SYMPOSIUM 2009
June 25-27, Graz

BROADBAND MULTIPOLE–MATCHED RENDERING: THE CURRENT STATE OF AFFAIRS

Jens Hannemann[1]

[1] University of Kentucky, Center for Visualization and Virtual Environments (j.hannemann@ieee.org)

Abstract: *Multipole–Matched Rendering (MMR) is a novel method to create three-dimensional sound fields in a listener's vicinity (the* sweet spot*). The main features of MMR are arbitrary speaker and sweet spot location as well as computational efficiency. It also requires less number of speakers than traditional approaches. Rendering is achieved by expanding both the sound field of a virtual source as well as those of the loudspeakers into multipoles located at the center of the sweet spot. The resulting error can be minimized using the Method of Moments (MoM) using either a Galerkin or pointmatching approach. This results in an usually overdetermined linear system of equations that can be solved in the least–squares sense using the pseudoinverse computed from a Singular–Value Decomposition (SVD). The SVD optimally matches the multipole expansion of the virtual source to that of the speakers — hence the term Multipole–Matched Rendering.*

This contribution reports the extension of the method to broadband signals and highlights some optimization strategies to reduce computational costs to arrive at a real–time implementation.

Key words: Spatial Sound, Multipole Expansion, Method of Moments, Singular–Value Decomposition

1 INTRODUCTION

Multipole–Matched Rendering (MMR) is emerging as an alternative to other sweet–spot solutions for rendering spatial sound such as Dolby Digital or Higher–Order Ambisonics (HOA) [1, 2]. It is closely related to HOA in that it also utilizes a spherical decomposition of the sound field. While HOA, however, uses all expansion coefficients up to a pre–determined order to satisfy a spatial sampling theorem, MMR uses the Singular–Value Decomposition (SVD) to optimally match the multipole expansions of a virtual source to those representing the speakers. As a consequence, MMR can typically use fewer speakers that do not need to be in a regular arrangement and still achieve acceptable perception. Rendering quality has been shown to be on-par with Wave–Field Synthesis (WFS) in perceptional experiments [3].

MMR has originally been derived for a single frequency [4]. Rendering of a narrowband signal was then achieved by computing the signal's Hilbert Transform, multiplying it with the computed speaker weights and playing back the real part of the computed signal on the actual speakers. This is briefly reviewed in section 2. Rendering broadband signals requires a full subband decomposition approach. Possible strategies for this and the current implementation are discussed in section 3. An optimization strategy for moving sources is presented in section 4.

2 MONOFREQUENT FORMULATION

The pressure field emanated from a single–frequency, omnidirectional sound source in free space is

$$p_s(\boldsymbol{r}_l, \boldsymbol{r}'_s) = A_s \frac{e^{-j\kappa|\boldsymbol{r}_l - \boldsymbol{r}'_s|}}{4\pi|\boldsymbol{r}_l - \boldsymbol{r}'_s|}\,. \tag{1}$$

$\boldsymbol{r}'_s$ and $\boldsymbol{r}_l$ are the source and listener locations, respectively. $\kappa = \frac{2\pi}{\lambda}$ is the wave number. Throughout this paper, a time factor of $e^{+j\omega t}$ is assumed. Source coordinates are denoted with a prime while observation coordinates are unprimed. In the following, a unit amplitude ($A_s = 1$) is assumed. This is the closed form expression for the free–space acoustic Green's function. The goal is to approximate the pressure field given by (1) (i.e. a virtual source) by an array of S loudspeakers of the same characteristic. The composite sound field of the loudspeaker array can be written as

$$p(\boldsymbol{r}_l, \boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_N) = \sum_{i=1}^{S} A_i \frac{e^{-j\kappa|\boldsymbol{r}_l - \boldsymbol{r}'_i|}}{4\pi|\boldsymbol{r}_l - \boldsymbol{r}'_i|}\,. \tag{2}$$

$\boldsymbol{r}'_i$ is the location of the i-th speaker and the A_i are unknown, complex *speaker weights* that need to be determined.

Due to the finite number of speakers, the approximation of the sound field according to (1) using (2) will generally contain an error e such that

$$p_s(\boldsymbol{r}_l, \boldsymbol{r}'_s) = p(\boldsymbol{r}_l, \boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_S) + e(\boldsymbol{r}_l, \boldsymbol{r}'_1, \ldots, \boldsymbol{r}'_S)\,. \tag{3}$$

Minimization of e according to to a suitable strategy and criterion will lead to a linear set of equations that can be

used to solve for the speaker weights A_i. To this end, the pressure fields of the virtual source and the speakers are expanded in terms of spherical multipoles. The spherical–multipole expansion of the free–space Greens function is [5, p. 259, eq. 8.22]

$$\frac{e^{-j\kappa|\boldsymbol{r}-\boldsymbol{r}'|}}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} = -j\kappa \sum_{n=0}^{\infty} \mathrm{j}_n(\kappa r_<)\mathrm{h}_n^{(2)}(\kappa r_>) \sum_{m=-n}^{+n} \mathrm{Y}_{n,m}(\vartheta,\varphi)\mathrm{Y}^*_{n,m}(\vartheta',\varphi') \,, \tag{4}$$

where the following definition holds for the radial coordinate:

$$r_> = \begin{cases} r & , r > r' \\ r' & , r < r' \end{cases} \quad \text{and} \quad r_< = \begin{cases} r & , r < r' \\ r' & , r > r' \end{cases} . \tag{5}$$

j_n is the spherical Bessel function ensuring regularity of the field at the origin and $\mathrm{h}_n^{(2)}$ is the Hankel function of the second kind, satisfying the Sommerfeld radiation condition for free space. Likewise, $\mathrm{Y}_{n,m}$ are spherical harmonics related to the Associated Legendre Functions of the First Kind, P_n^m by:

$$\mathrm{Y}_{n,m}(\vartheta,\varphi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}\mathrm{P}_n^m(\cos\vartheta)e^{jm\varphi} \,. \tag{6}$$

The spherical harmonics are an orthogonal function system, i.e.

$$\int_0^{\pi}\int_0^{2\pi} Y_{n,m}(\vartheta,\varphi)Y^*_{n',m'}(\vartheta,\varphi)\sin\vartheta \, d\vartheta \, d\varphi = \delta_{n,n'}\delta_{m,m'} \,, \tag{7}$$

where $\delta_{i,j}$ is the Kronecker delta. This orthogonality will be used to efficiently compute the speaker weights A_i.

To this end, the closed–form expressions in (3) are replaced by the corresponding multipole expansions according to (4). To exploit orthogonality, the error is minimized on average over a sphere around the listener (radius r_l). It is further assumed that all virtual sources are located outside the listener's head. This means that $r_> = r'$ and $r_< = r = r_l$ hold for all virtual and physical sources. Equation (3) thus becomes

$$\sum_{n=0}^{\infty} \mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_s) \sum_{m=-n}^{+n} \mathrm{Y}_{n,m}(\vartheta_l,\varphi_l)\mathrm{Y}^*_{n,m}(\vartheta'_s,\varphi'_s) = \sum_{i=1}^{S} A_i \sum_{n=0}^{\infty} \mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_i) \sum_{m=-n}^{+n} \mathrm{Y}_{n,m}(\vartheta_l,\varphi_l)\mathrm{Y}^*_{n,m}(\vartheta'_i,\varphi'_i) + e(\boldsymbol{r}_l,\boldsymbol{r}'_1,\ldots,\boldsymbol{r}'_S) \,. \tag{8}$$

The so–called Galerkin approach for the Method of Moments mandates that the average error on the sphere around the listener's head becomes zero, i.e. that

$$\int_0^{\pi}\int_0^{2\pi} e(\boldsymbol{r}_l,\boldsymbol{r}'_1,\ldots,\boldsymbol{r}'_N)\mathrm{Y}^*_{n,m}(\vartheta_l,\varphi_l)r_l\sin\vartheta_l \, d\vartheta_l \, d\varphi_l \stackrel{!}{=} 0 \,. \tag{9}$$

The Kronecker deltas from the orthogonality relation (7) filter out a single term from the double sum over m and n in (8). Using the first N_R radial modes, this leads to a system of N_R^2 linear equations. For any given combination of n and m, the equation becomes

$$\mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_s)Y^*_{n,m}(\vartheta'_s,\varphi'_s) = \sum_{i=1}^{S} A_i \mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_i)\mathrm{Y}^*_{n,m}(\vartheta'_i,\varphi'_i) \,. \tag{10}$$

The system of equations can be rewritten as

$$[C_{j,i}]\,[A_i] = [B_j] \,, \tag{11}$$

where j is related to n and m by $j = n^2 + n + m$, where $n = 0,\ldots,N_R$ and $m = -n,\ldots,+n$. Let $N = N_R^2$. Then C is an $N \times S$ matrix, $\boldsymbol{A}$ is an $S \times 1$ column vector and $\boldsymbol{B}$ is an $N \times 1$ column vector. The matrix entries are

$$C_{j,i} = \mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_i)\mathrm{Y}^*_{n,m}(\vartheta'_i,\varphi'_i) \,. \tag{12}$$

The A_i are the speaker weights and the right–hand side elements are

$$B_j = \mathrm{j}_n(\kappa r_l)\mathrm{h}_n^{(2)}(\kappa r'_s)\mathrm{Y}^*_{n,m}(\vartheta'_s,\varphi'_s) \,. \tag{13}$$

The system of equations (11) is typically overdetermined, although a very large number of speakers can also lead to an underdetermined system. Such a system can be solved optimally in the least–squares sense using an SVD. The SVD of matrix C is given by [6]:

$$C = U\Sigma V^H \,, \tag{14}$$

where the superscript H denotes the Hermitian. U is $N \times N$, Σ is $N \times S$, and V^H is $S \times S$. The pseudo–inverse C^+ is then of dimension $S \times N$ and can be computed by

$$C^+ = V\Sigma^+U^H \,, \tag{15}$$

where Σ^+ is obtained by replacing the non–zero singular values in Σ by their respective inverse. Besides being the optimal solution, the singular values in the pseudo–inverse can be manipulated to numerically stabilize or accelerate the solution. In a broadband context, the singular values can also be used to achieve equalization to offset or minimize coloration.

The matrix elements in equation (12) only depend on the speaker and listener location. This means that for stationary speakers and listener, rendering of a moving source is efficiently achieved by a simple repeated computation of the vector $\boldsymbol{B}$ and a subsequent matrix–vector multiplication

$$\boldsymbol{A} = C^+\boldsymbol{B} \,. \tag{16}$$

3 BROADBAND EXTENSION

The first step in extending MMR to broadband signals is to choose an appropriate number of frequency bins, N_f, for the subband decomposition. Then, for each frequency bin, the pseudo–inverse and, if the source has moved, the right–hand side according to (12), (15), and (13) are computed. Subseqent multiplication of C^+ and $\boldsymbol{B}$ yields S source–to–speaker transfer functions $A_i(k)$ of size N_f. At this point, two different paths can be taken. If the main processing of the audio input is to be done in the frequency domain, the input signal will be decomposed using an FFT. The resulting input signal spectrum will then be multiplied with the S transfer functions, subjected to an inverse FFT and output to the speakers. An alternative path is to compute the inverse FFT of the source–to–speaker transfer functions, which yields S source–to–speaker impulse responses

$$a_i(n) = \sum_{k=0}^{N_f-1} A_i(k) e^{j\frac{2\pi}{N_f}kn} . \tag{17}$$

The audio input $x(n)$ is subsequently convolved with those impulse responses

$$y_i(n) = x(n) * a_i(n) \sum_{m=0}^{N_f-1} x(n) a_i(n-m) \tag{18}$$

and then output to the speakers. This is the approach that has been used in [3].

4 OPTIMIZING PERFORMANCE

For stationary sources, performance optimization is a non–issue, as the main processing loop in this case simply consists of the S convolutions. An engine like BruteFIR [7] (also used in [3]) easily achieves realtime performance for more than 50 channels. For a moving source, the repeated computation of $\boldsymbol{B}$ according to equation (13) becomes the dominant part. The computationally most expensive part is the evaluation of the spherical cylinder functions, which must be computed via a backwards recurrence that is very compute–intensive, especially for higher orders.

In the current implementation, the path for a moving source is described using a parametrized curve $\boldsymbol{p}(s)$ in 3D space, along with the source's velocity v along the curve. If s is the natural parameter (i.e. the parameter value is equal to the distance traveled from the starting point along the curve), then the position $\boldsymbol{s}$ of a source can be computed from the time t as

$$\boldsymbol{s}(t) = \boldsymbol{p}(vt) . \tag{19}$$

In the simplest case, the path is sampled at the same rate as the audio. For a typical 44.1kHz sample rate and a linear path with a speed of $15\frac{\mathrm{m}}{\mathrm{s}}$, this translates into a distance of 0.3mm traveled between samples. While realtime operation has been achieved for up to 10 channels at this full spatial resolution using multithreading on a quad–core 2.8GHz Xeon system, full spatial sampling is clearly unnecessary. Using the natural parameter on a curve makes it easy to determine the distance a source has traveled along its path, so an obvious optimization strategy is to recompute (13) only after the source has moved farther than a threshold distance from the last sample location. However, even for a modest reduction in spatial resolution, clicks are audible in the rendered sound. The reason is that the corresponding impulse responses are sufficiently different to create audible discontinuities when the source reaches the next spatial sample point. The discontinuities can be avoided by using linear interpolation of the impulse responses between sample points. Using this technique, the update interval can be increased even further while still maintaining a subjectively correct impression of source localization and movement. Figure 1a illustrates the discontinuities generated in the difference between a signals with a spatial update of 1cm (interpolated) and a spatial update of 10cm (not interpolated). Figure 1 shows the difference between two signals for update intervals of 1cm and 100cm (both interpolated).

Spatial Sampling	Interpolation	CPUs	Runtime
1cm	yes	70	11.082min
10cm	no	70	2.755min
100cm	yes	28	1.587min

Table 1: Runtimes to render 5s worth of audio data, using different spatial update intervals and interpolation methods. The results have been computed using octave on a computer cluster using the indicated number of CPUs (2GHz Intel Xeon). The number of speakers was $S = 40$, the number of radial modes was $N_R = 14$, and 258 frequency bins have been used.

5 SUMMARY AND OUTLOOK

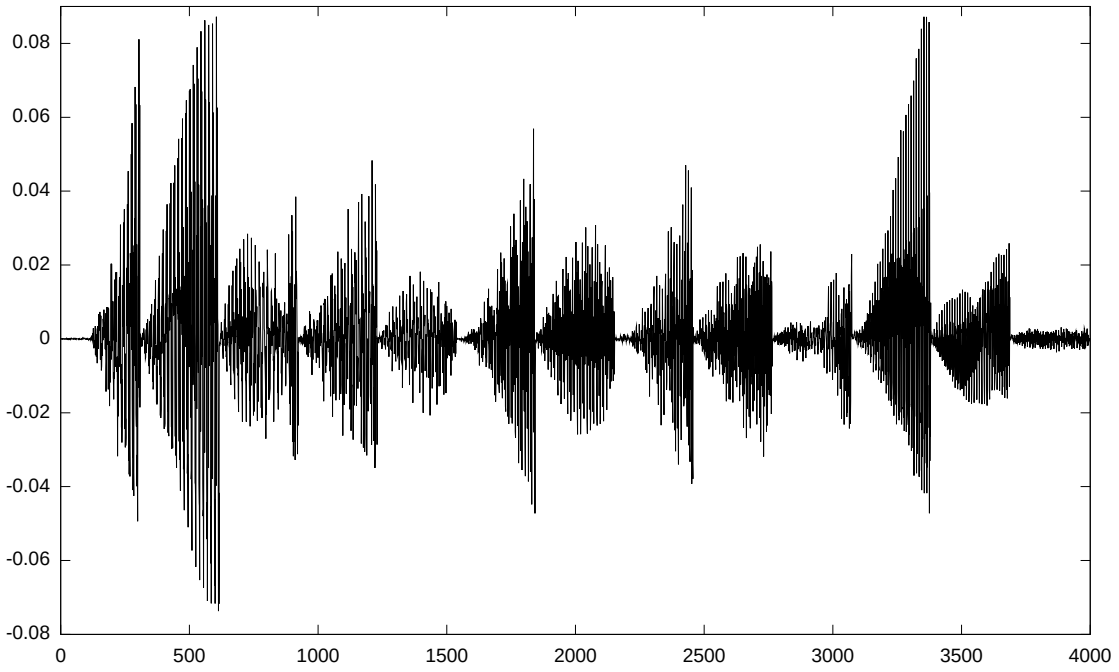
MMR has successfully been extended to broadband signals. Decoupling the spatial from the temporal sampling rate results in significant speedups while maintaining rendering quality. Currently, an implementation of the method that will achieve realtime operation on an SMP system is underway. Future work will concentrate on more formal evaluation of the impact of the number of frequency bins as well as studying coloration.

REFERENCES

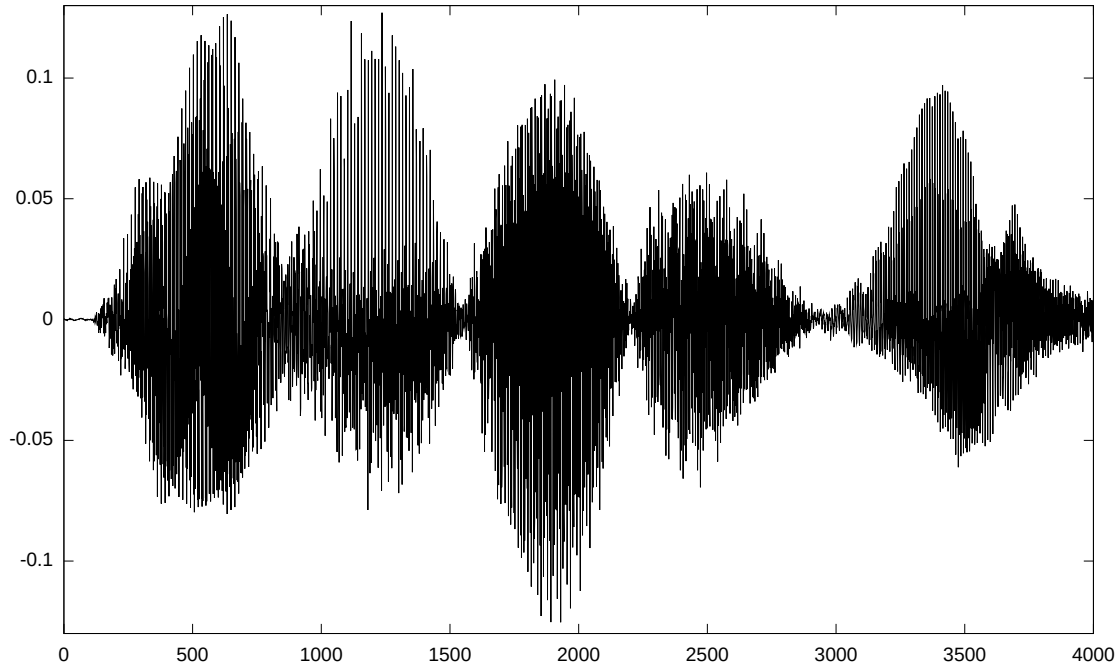
[1] Jérôme Daniel, Rozenn Nicol, and Sébastien Moreau. Further investigations of high-order ambisonics and wavefield synthesis for holophonic sound imaging. In *114th Convention of the AES*. Audio Engineering Society, March 2003.

[2] Jérôme Daniel and Sébastien Moreau. Further study of sound field coding with higher order ambisonics. In *116th Convention of the AES*. Audio Engineering Society, May 2004.

[3] Jens Hannemann, Christopher Leedy, Kevin Donohue, Sascha Spors, and Alexander Raake. A comparative



(a) Difference between signals rendered using a 1cm (interpolated) and 10cm (non–interpolated) update interval, respectively. As the source moves away from the last valid sample point, the differences in the impulse responses increase. The maximum absolute amplitude error is about -12dB. The discontinuities occur as the source reaches the next valid sample point and the error suddenly snaps back to zero. This is clearly audible as clicks in the rendering.



(b) Difference between signals of a moving source sampled using a 1cm and 100cm update interval, both interpolated. The maximum absolute amplitude error is still about -12dB, however, no clicking is present in the rendering any more as the transition between the two error–free sampling points has been smoothed by the interpolation. Subjectively, localization and impression of movement is not impeded.

Figure 1: Difference between a typical channel of the signals of a moving source sampled at various spatial resolutions with and without linear interpolation. Linear interpolation allows for much greater spatial sampling intervals while still maintaining rendering quality, resulting in large execution time speedups.

study of the perceptional quality of wave field synthesis versus multipole matching for audio rendering. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 397–400. IEEE, April 2008.

[4] Jens Hannemann and Kevin Donohue. Virtual sound source rendering using a multipole–expansion and method–of–moments approach. *Journal of the Audio Engineering Society*, 56(6):473–481, June 2008.

[5] Earl George Williams. *Fourier Acoustics*. Academic Press, New York, NY, 1999.

[6] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[7] Anders Torger, June 2009. URL `http://www.ludd.luth.se/~torger/brutefir.html`.