Proceedings of the International Symposium on Room Acoustics, ISRA 2010

29–31 August 2010, Melbourne, Australia

Soundfield synthesis with spatio-temporal compensation

D. Sen (1), S. Wang (1) and Q. Meng (1)

(1) School of EE&T, University of New South Wales, Sydney, Australia

PACS: 43.60.Fg, 43.60.Sx, 43.55.Br

ABSTRACT

Practical soundfield synthesis systems are required to recreate the original soundfield at a multitude of diverse arbitrary acoustic environments. For accurate reconstruction, it is essential that the local acoustic characteristics are equalized at the playback venue. This requires physical measurement and monitoring of the local spatio-temporal characteristics. In this paper, we present an implementation for spherical harmonic based soundfield synthesis systems which uses previously published methods for measuring the local spatio-temporal response. The multipoint measurement uses an amplitude modulated log-swept stimuli which is subsequently incorporated as time domain inverse filters into the spherical-harmonic based synthesis framework. The time-domain inversion of the impulse responses are non-trivial in practice due to their mixed-phase characteristics. The complete methodology accounts for the spatio-temporal response of each loudspeaker (for arbitrary loudspeakers and playback environments) as well as local room acoustics. The recreated soundfield is evaluated by asking a pool of listeners to perform Ref/A-B tests. The reference stimuli is a binaural recording of the same soundfield using a B&K Head and Torso Simulator. The stimuli was presented to the listeners using headphones. The A & B candidates are the equalized and non-equalized soundfield synthesized using 25 loudspeakers in a non-anechoic environment.

INTRODUCTION

Soundfield synthesis systems are often implemented in anechoic laboratory conditions. For practical deployment to consumers however, the original soundfield has to be recreated at arbitrary non-anechoic playback venues. For such applications, it is essential that the synthesis algorithm accounts for the local acoustic characteristics at the playback venue. To avoid the requirement for an explicit measurement, loudspeaker radiation models have been used in [1]. However, no matter how accurate the modeling, it is impossible to predict the acoustic landscape at arbitrary locations. Other techniques have been proposed in [2–4] but have often proven problematic due to their inability to delineate the linear response from non-linear distortions as well as not being conducive for use in soundfield synthesis algorithms. In [5], a 2D modal approach is shown in simulation to outperform the multipoint approach by 5 dB across the frequency range. The spatio-temporal response was calculated using the image method. In the current work, we use actual multipoint measurements in 3D to calculate the room impulse response. We find that such measurements are more conducive to a time-domain inverse filtering approach for equalization purposes due to their mixed-phase characteristics [6].

In this paper, a complete signal processing frame-work is presented, whereby the spatio-temporal response measured using amplitude modulated exponential sine swept stimuli [7] is incorporated into the soundfield synthesis system using a time-domain inverse filtering approach. This time domain approach is mathematically equivalent to a full-matrix frequency domain inversion (least squares) approach - however with computational complexity benefits. The purpose of the work is to empirically test the validity of the spatio-temporal measurement, explore the feasibility and methodology of finding its inverse (usually complicated by their mixed-phase characteristics), the practical feasibility of the complete soundfield acquisition/synthesis frame-work and the perceptual performance of the complete system.

The frame-work from soundfield acquisition to reproduction - incorporating the required equalization to account for spatio-temporal response of the local (room and loudspeaker) environment is shown in Figure 1. The diagram depicts the acquisition of the original soundfield using a microphone array "Mic Array 1", $p_i^o[n], 0 < i < N$, (where n is the discrete time index, and N is the number of microphones in the array). The microphone signals are subsequently represented as coefficients of an orthogonal projection into Spherical Harmonics basis functions (detailed in the next Section). For the unequalized case, this representation is decoded directly into loudspeaker feeds for the known loudspeaker positions (in the local playback environment). This path is shown in Figure 1 as using "Decoding Matrix 1". The decoding matrix in this case is a function of the loudspeaker coordinates.

The equalized system is intended to replicate original soundfield in a closed 3D area of the local environment. To that end, we can use a second decoding matrix ("Decoding Matrix 2" in Figure 1) to calculate the pressure at multiple points within that area. The decoding matrix in this case is a function of the co-ordinates of these multiple points. In practice, these points are the locations of microphones of a second microphone array ("Mic Array 2") used to measure the spatio-temporal response of the local environment. We call these the desired pressure signal, $p_i^d[n], 0 < i < L$ (where L is the total number of microphones in "Mic Array 2"). If the acoustic transfer function, h_{ij}, from each loudspeaker (indexed by j) to each microphone in the array is known, then it is possible to reproduce the desired pressure at each of these points by calculating speaker feeds

which are given by inverse filtering each $p_i^d[n], 0 < i < L$ with h_{ij}^{-1}. The hypothesis within this work is that by ensuring the pressure at the microphone locations of "Mic Array 2" are exactly the same as that of the original soundfield, the perceived sound *in the vicinity* of "Mic Array 2" will closely resemble that of the original soundfield.

SPATIO-TEMPORAL COMPENSATION IN SOUNDFIELD SYNTHESIS SYSTEMS

In the class of Spherical Harmonic based soundfield analysis and synthesis systems [8] (which includes the well known Ambisonics systems [3, 9] and beamforming applications [4]) the scalar pressure field is represented by a series expansion as follows:

$$p(r,\theta,\phi,k) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_n^m(k) j_n(kr) Y_n^m(\theta,\phi) \qquad (1)$$

where r,θ,ϕ are positional variables of radius, elevation angle and azimuth angle, k is the spatial frequency, $Y_n^m()$ are the Spherical harmonics of degree m and order n and $j_n()$ is the n-th order spherical Bessel function of the first kind. Further, we assume time-harmonic representations of all signals and the time variation of $e^{j\omega t}$ is implicit and not mentioned hereafter. The series coefficients of this expansion $A_n^m(k)$ are a complete description of the soundfield present at the recording location. The series expansion in Equation 1 can be truncated at sufficiently large $n = N$ with minimal error due to the nature of spherical Bessel functions which decay rapidly at higher orders and high values of kr. The extraction of these coefficients is the field of soundfield analysis and a good review of these techniques are given in [8].

In the following we assume that a suitable technique (such as in [10]) has been used to record the soundfield (at an arbitrary location) and that the soundfield is represented by a set of coefficients $A_n^m(k)$ as per Equation 1. This enables the focus of this paper which is the re-synthesis and the framework to compensate for the spatio-temporal response at the playback (local) environment. For this purpose, lets assume that we can sample the pressure field at $L >= (N+1)^2$ arbitrary positions, $\mathbf{r_i} = (r_i,\theta_i,\phi_i), 0 \le i < L$, in the playback environment. Equation 1 can then be likened to a matrix operation as follows:

$$\begin{pmatrix} p(\mathbf{r_0},k) \\ p(\mathbf{r_1}k) \\ \vdots \\ p(\mathbf{r_{L-1}},k) \end{pmatrix} = [D_{l,n}] \begin{pmatrix} A_0^0(k) \\ A_1^{-1}(k) \\ A_n^m(k) \\ \vdots \\ A_N^N(k) \end{pmatrix} \qquad (2)$$

where, the vector on the left hand side of the equation represents the pressure $p(\mathbf{r_i})$ at the locations $\mathbf{r_i}$, at a given spatial frequency k and D is a matrix of size $L \times (N+1)^2$ whose elements, are given by:

$$D_{l,n} = j_n(kr_l) Y_n^m(\theta_l,\phi_l). \qquad (3)$$

The pressure vector in Equation 2 can be interpreted as the ideal pressure that is to be reproduced at the spatial positions $\mathbf{r_i}$ to create the soundfield represented by the coefficients $A_n^m(k)$. This formulation while similar to Higher Order Ambisonics (HOA) has some subtle implementation differences. It is used in [1, 8] for example. The most conspicuous difference, that of the inversion of the spherical-harmonic matrix which in this method is carried out at the encoder rather than the decoder (for HOA). This can however be mathematically shown to be equivalent. The problem in the context of soundfield synthesis, however, is to compute the loudspeaker feeds required to produce this pressure at the locations $\mathbf{r_i}$. The pressure at these discrete locations, $\mathbf{r_i}$, will be influenced by each of the outputs from the loudspeakers, their radiation patterns, as well the acoustic environment (reflections, absorption, etc) at the playback environment.

For each loudspeaker, located at $\mathbf{r_j}, 0 \le j < M$, (assuming a total of M loudspeakers) the local spatio-temporal response at the arbitrary positions $\mathbf{r_i} = (r_i,\theta_i,\phi_i), 0 \le i < L$ is given by $H_{ij}(\mathbf{r_j},\mathbf{r_i},k)$ where H_{ij} relates the speaker feed $s_j(k)$ to the pressure $p_i(k)$ at a location $\mathbf{r_i}$ as follows:

$$H_{ij}(\mathbf{r_j},\mathbf{r_i},k) s_j(k) = p_i(k). \qquad (4)$$

The above equation states that the pressure at position $\mathbf{r_i}$ is composed of M convolution processes; between each speaker feed and its corresponding spatio-temporal response at the location $\mathbf{r_i}$. Correspondingly, this means that the speaker feeds can be computed by the filtering of $p_i(k)$ by the M inverse filters h_{ij}^{-1}, i.e.

$$s_j(t) = p_i(t) * h_{ij}^{-1}(\mathbf{r_j},\mathbf{r_i},t). \qquad (5)$$

The process thus requires the measurement of $L \times M$ spatio-temporal responses at the playback environment, followed by the computation of the corresponding inverse filters h_{ij}^{-1}. The latter is complicated by the fact that the spatio-temporal responses are not necessarily minimum phase [6, 11] making the computation of the inverse filters non-trivial. The next section describes the spatio-temporal measurement process using exponential sine swept stimuli. This was presented earlier [7] and summarised here for context and completeness. Subsequent section describes the computation of the set of inverse filters. The overall system was tested and validated by (i) recording a soundfield using both a 32 channel spherical microphone array (Eigenmike from MHAcoustics) and a binaural recording (using a B&K 4100D Head and Torso Simulator) (ii) reconstructing the soundfield in a reverberant environment. For comparison, this was done both with and without the equalization method described above (iii) conducting REF/A/B subjective tests where the listeners were asked to identify the soundfield recreation that best matched the binaural recording (the A/B candidates being the equalized and unequalized synthesis and the reference the binaural recording played back through Sennheiser HD650 headphones). These results are presented in the last section.

METHODOLOGY

Impulse response measurement using exponentially sine swept stimuli [1]

This section describes the measurement of the spatio-temporal response $H_{ij}(k)$ for each loudspeaker k ($0 \le k < M$) and L arbitrary positions $\mathbf{r_i}$. The arbitrary positions are typically the positions of microphone modules in some geometrical array configuration.

The generic form of an exponentially swept sine signal is given by:

[1] The discussion of this section was provided in [7]. It is reproduced here for completeness.

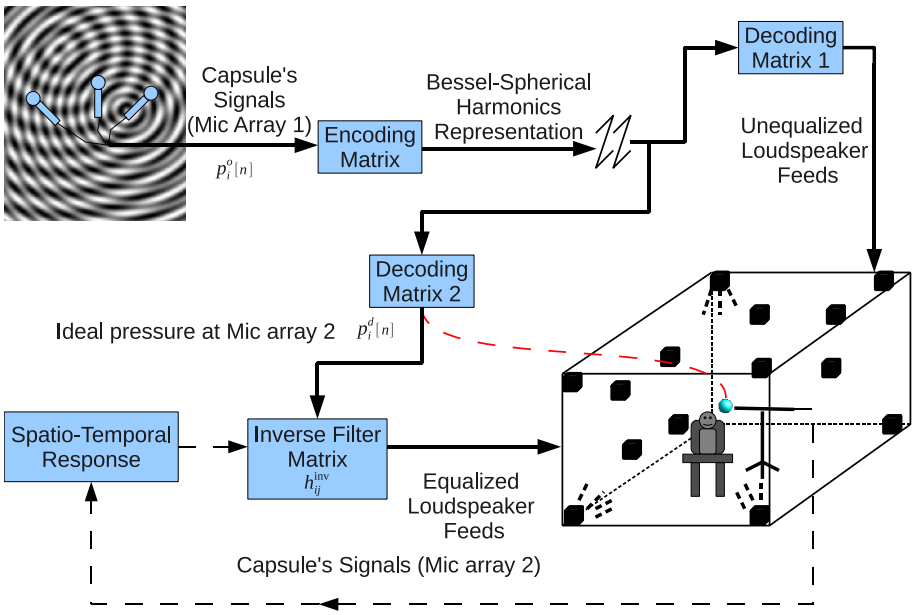


Figure 1: Framework of sound field synthesis system with spatio-temporal compensation. The red dashed line indicates that the output of Decoding Matrix 2 produces the ideal pressure at precise points (in practice the microphone positions of Mic Array 2) within a small area of the local listening environment.

$$s(t) = \sin[K \cdot (e^{\frac{t}{L}} - 1)], \tag{6}$$

where,

$$K = \frac{T \cdot \omega_1}{\ln(\frac{\omega_2}{\omega_1})}, L = \frac{T}{\ln(\frac{\omega_2}{\omega_1})} \tag{7}$$

and ω_1 and ω_2 are the lower and higher extremities frequency range of the measurement, respectively. T is the duration of the stimuli, in seconds. The instantaneous frequency $\omega(t)$ is then given by:

$$\omega(t) = \frac{d[K \cdot (e^{\frac{t}{L}} - 1)]}{dt} = \frac{K}{L} \cdot e^{\frac{t}{L}} \tag{8}$$

The time variation of the energy of a sinusoidal signal with instantaneously varying frequency, can be shown to be inversely proportional to the rate of change in frequency. Thus, the energy as a function of time $E(t)$ is given by:

$$E(t) \propto \frac{1}{\omega'(t)} = \frac{L^2}{K} \cdot e^{-\frac{t}{L}}. \tag{9}$$

Thus, as a function of frequency, the energy density $E(\omega)$, is given by:

$$E(j\omega) = \frac{\alpha L^2}{K} \cdot \frac{1}{L + j\omega}, \tag{10}$$

where α is a constant of proportionality. This represents a -3dB/octave slope which would not be present in a linear-swept sinusoid, where the time rate of change of frequency, $\omega'(t)$ is constant. We note here that the time-reversed version of the stimuli has exactly the same energy distribution. Figure 2 shows this energy distribution for a stimulus with low (starting) frequency $\omega_1/2\pi = 20$ Hz, high (ending) frequency $\omega_2/2\pi =$ 21,000 Hz and time duration $T = 10$ seconds. Due to the sudden switch-on at the beginning and switch-off at the end, unwanted ripple appears at the extremities of the spectrum [12]. Half-windows (tapered cosine) are used to smooth this impact.

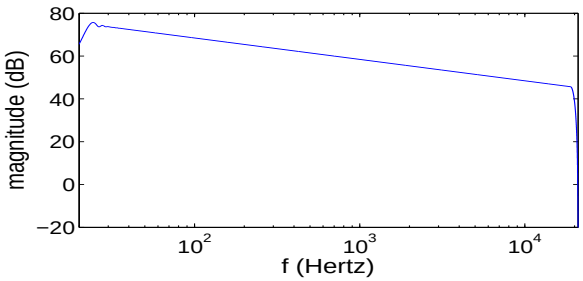


Figure 2: Magnitude response of exponential sine sweep

The methodology in [13] involves exciting the system under measurement with the stimuli in Eq. 6, and convolving the output with the time reversed version of Eq. 6. The output of this convolution can be used to extract the impulse response as long as the slope of 6 dB/octave (due to the sequential 3dB/octave effect of convolving with the stimuli and its time-reverse) is accounted for. In [14], a post-processing strategy whereby the time-reversed stimuli is amplitude modulated to produce a modified signal (inverse filter) that has an effective slope of +3dB/octave. To produce this, a modulating signal of the form given by Eq. 11 is required.

$$m(t) = \frac{A}{\omega(t)} = A(\frac{K}{L} \cdot e^{\frac{t}{L}})^{-1} \tag{11}$$

Arbitrarily setting the value of $m(t) = 1$ at $t = 0$, we get $A = \omega_1$.

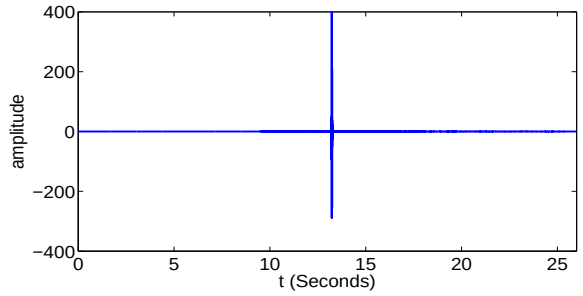
We modify the approach slightly such that instead of the post-processing strategy described above, we propose a direct amplitude modulation of the input sine swept signal. This pre-processing strategy involves the use of a modulating signal with a slope of +3dB/octave to account for the -3 dB/octave slope of the log-swept sine stimuli. The modified signal with

a flat spectrum can be used as the input stimuli to excite the system under measurement. No post-processing is required beyond the convolution by the time reverse of the proposed modified input signal. The general form of the new modulating function with a +3dB/octave is given by:

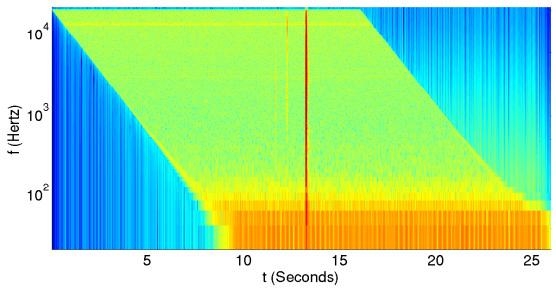
$$n(t) = B \cdot \sqrt{\omega(t)} \tag{12}$$

Arbitrarily setting the value of $n(t) = 1$ at $t = 0$, we get $B = \frac{1}{\sqrt{\omega_1}}$.

The deconvolved response and its spectrogram using the pre-processing strategy using an amplitude modulated input stimuli are shown in Fig. 3. As the spectrogram reveals, the harmonic distortions are represented by the vertical lines parallel and to the left (on the time axis) of the actual linear response which starts at about 13 seconds on the spectrogram. The harmonic distortions are located at precise anticipatory times [15] on the time axis. The almost 2 second wide spacing between the linear impulse response and the closest harmonic distortion represents a good separation between linear response and non-linear distortion. The linear response and harmonic distortions are well separated. Due to our arbitrary choice of $n(t) = 1$ at $t = 0$, the SNR at extremely low frequencies is not as high as the impulse response extracted by the post-processing method. This can be improved by selecting a higher value of $n(t)$ at $t = 0$.



(a) impulse response



(b) spectrogram

Figure 3: Measured impulse response (top) and spectrogram (bottom) using the pre-processing method. The 26-second long response is explained by the convolution between the 10 second (time reversed) stimuli and the 16 second recorded signal

Inverse filter calculation and use

The above section describes an accurate method of calculating the spatio-temporal response of arbitrary rooms. However, to calculate the speaker feeds, the inverse filters corresponding to these responses are required (as per Eq. 5). The invertibility of the responses have been studied [6, 11, 16, 17] and complicated by the fact that the responses are usually mixed phase. Least squares techniques are also hampered by having to find an optimum delay [16] in the minimization criteria.

Our solution to the problem has been to treat the minimum and maximum phase components as causal and non-causal components in the inverse response. The non-causal response can be accounted for by prepending an appropriate number of zeros to the spatio-temporal response before calculating their inverse response using a DFT. This is akin to introducing a delay to account for the non-causal component. The complete process is thus as follows:

Step 1: Synchronize the set of $L \times M$ spatio-temporal response sequences $h_{ij}[n]$ keeping a record of their individual delays T_{ij}.

Step 2: Select appropriate length of the response, ensuring the last points coincide with noise thresholds of the response.

Step 3: Prepend zeros to the responses to create a new sequence $h'_{ij}[n]$:

$$h'_{ij}[n] = \begin{cases} 0.0 & 0 \leq n < A \\ h_{ij}[n] & n \geq A \end{cases} \tag{13}$$

Step 4: Calculate the DFT of the sequence $DFT\{h'_{ij}[n]\}$.

Step 5: Calculate the inverse DFT of the reciprocal of the sequence calculated in Step 4.

Step 6: Convolve the sequence obtained in Step 5 with $\delta[n - T_{ij}]$ obtained in Step 1. The output of this final step are the inverse filters $h_{ij}^{-1}[n]$. An optional filter which attempts to correct the unrealistic gains around $f_s/2$ [17] is carried out. These inaccuracies can clearly be seen in the low gains around $f_s/2$ of the forward filter $H_{ij}(k)$ in Figure 5. Attempting to correct for this low gain will result in the excessive and inaccurate boost at the high frequencies reported in [17].

Figures 4 - 6 show an example of a measured spatio temporal response in the time and frequency domain and its corresponding inverse calculated using the above steps. Figure 7 shows the convolution of $h_{ij} * h_{ij}^{-1}$ which closely approximates an unit impulse (delayed by T_{ij}). Figure 7 shows the accuracy of the methodology described above. We verified that all 800 such plots produced a co-incident impulse - indicating that both the phase and magnitude response of the local environment had been accounted for.

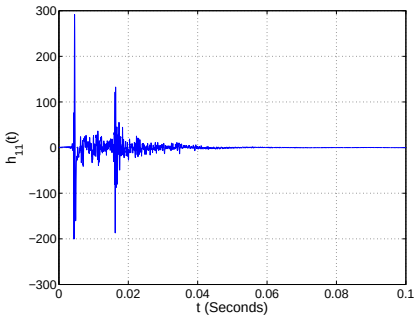


Figure 4: Spatio-temporal response example $h_{ij}(t)$ as a function of time.

Each speaker feed $s_i[n], 0 < i < M$ is computed as per the following equation:

$$s_i[n] = \frac{1}{N} \sum_{j=1}^{N} p_j[n] * h_{ij}^{-1}[n], \tag{14}$$

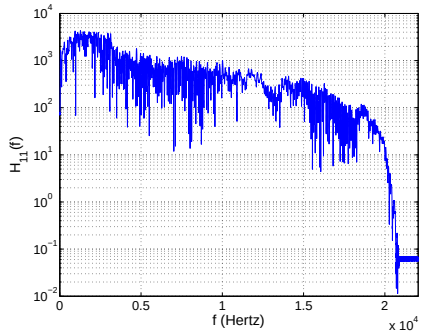


Figure 5: Spatio-temporal response example $H_{ij}(f)$ as a function of frequency.

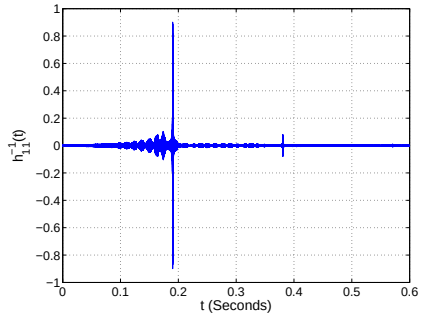


Figure 6: Inverse response example $h_{ij}^{-1}(t)$ calculated using Steps 1-6.

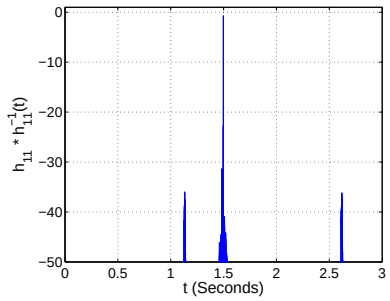


Figure 7: The convolution of $h_{11}[n]$ with $h_{11}^{-1}[n]$. The artifacts to the left of the main peak are approximately 40 dB below the main peak.

where N is the total number of microphones, M is the total number of loudspeakers and $p_j[n]$ represents the pressure calculated at each microphone position of the second microphone array (Mic Array 2) using Equation 2. The time domain approach described above can be shown to be mathematically equivalent to a full matrix inversion in the frequency domain. A frequency domain approach is however computationally prohibitive given the length of the inverse filter (as evidenced by this study) especially when considering that an overlap add procedure would require DFT sizes of twice this length. The results in the next section shows that the reproduced soundfield is perceptually closer to the original compared to a non-equalized synthesis.

RESULTS

The soundfield created by multiple orchestral instruments (playing the Nutcracker opera) across a diverse area of a large room (concreted and carpeted) of size 12.0 x 9.0 x 3.3 m was recorded in two different positions using a 32 channel mhAcoustics spherical microphone array and a B&K 4100D Head and Torso Simulator. The latter is a binaural recording that serves as the reference stimuli in the final Ref/A/B testing. The recording from the microphone array is converted to the $A_n^m(k)$ coefficients as described in [10]. These coefficients are decoded using two different methods. The first directly computes the pressure at 25 loudspeaker positions according to Equation 2 and uses these directly as the speaker feeds. This represents the unequalized stimuli. The 25 Genelec 8030A loudspeakers (driven by an RME M32DA MADI device) were situated in a room of size 5.44 x 3.66 x 2.54 m.

The second method decodes the $A_n^m(k)$ coefficients into the microphone positions of an arbitrary microphone array. Due to the availability and convenience, we used the same 32 channel MHAcoustics spherical microphone array (as for the recording) for this purpose. However, it is envisaged that any other microphone array would suffice. The speaker feeds were calculated by the inverse filtering approach described by Equation 5.

The spatio-temporal measurements h_{ij} were carried out from each of the 25 loudspeakers to each of the 32 microphones. The recordings were carried out at a sampling rate of 44.1 kHz with 24 bits/sample resolution and the measurements carried out as described in a previous section.

Ref/A/B subjective testing was carried out to evaluate the effectiveness of the equalization method. Using double blind testing, three subjects were asked to identify the stimuli (A or B) which most resembled the reverberant conditions of the reference binaural recording. The reference stimuli was played through a pair of Sennheiser HD 650 headphones. The graphical user interface presented to the listeners is shown in Fig 8. The stimuli consisted of the Nutcracker opera recorded at two different positions of the large room (described above). Three expert listeners were used and results as shown in Table 1 were unanimously in favor of the equalized synthesis.

Table 1: Result of the Ref /A /B test.

	Equalized	Unequalized
Set_1	100%	0%
Set_2	100%	0%

CONCLUSION

We have described a complete framework for synthesizing immsersive soundfields in environments that are not anechoic in nature. This is a realistic scenario for practical deployment into

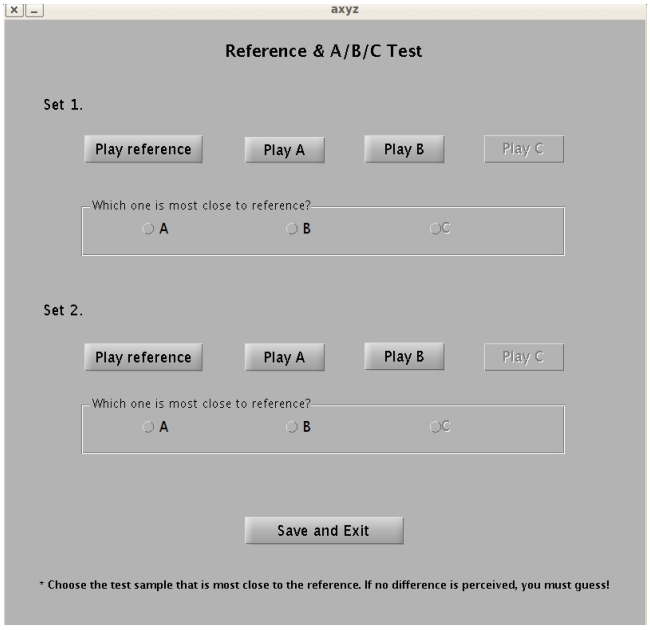


Figure 8: Graphical user interface for the Ref/A/B test

living rooms and cars. In particular it demonstrates the viability of recreating soundfields in environments away from the ideal anechoic chambers of acoustic laboratories. The spatio-temporal measurement technique produces highly accurate results that are devoid of non-linear distortions. The technique of compensating for the local acoustic environment and speaker characteristics can be used in all soundfield analysis and synthesis systems based on spherical harmonics and requires a one time measurement using a set of microphones at the playback location.

REFERENCES

1 A.Laborie R.Bruno and S.Montoya. Reproducing multichannel sound on any speaker layout. *118th AES Convention, Barcelona, Spain*, May 2005.
2 T.D. Abhayapala and D.B. Ward. Theory and design of high order soundfield microphones using spherical microphone array. *Proc. ICASSP, Orlando, Florida*, May 2002.
3 J.S. Bamford. An analysis of ambisonic sound systems of first and second order. *ME Thesis, University of Waterloo*, 1995.
4 J. Meyer and G. Elko. A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield. *Proc. ICASSP*, 2002.
5 T. Betlehem and T.D. Abhayapala. Theory and design of sound field reproduction in reverberant rooms. *The Journal of the Acoustical Society of America*, 117:2100, 2005.
6 P.D. Hatziantoniou and J.N. Mourjopoulos. Errors in real-time room acoustics dereverberation. *J. Audio Eng Soc.*, 52(9), 2004.
7 Q.Meng D.Sen S.Wang and L. Hayes. Impulse response measurement with sine sweeps and amplitude modulation schemes. *Proc. IEEE International Conference on Signal Processing and Communication Systems*, 2008.
8 M.A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. *J. Audio Eng. Soc.*, 53(11): 1004–1025, 2005.
9 P. Fellgetti. Ambisonics. part one. general system description. *Studio Sound*, 17(8), 1975.
10 D. Sen S. Wang and A. Deffrasness. Psychoacoustically motivated frequency dependent tikhonov regularization for soundfield parameterization. *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2010.
11 S.T. Neely and J.B. Allen. Invertibility of a room impulse response. *J. Acoust. Soc. America*, 66(1), 1979.
12 S.Muller and P. Massarini. Transfer function measurement with sweeps. *J. Audio Eng. Soc.*, 49:443–471, June 2001.
13 A. Farina. Simultaneous measurement of impulse response and distortion with a swept sine technique. *Presented at the 108th AES Convention, Paris, France*, 2000.
14 A. Farina. Advancements in impulse response measurements by sine sweeps. *Presented at the 122nd AES Convention, Vienna, Austria*, May 2007.
15 M.A. Poletti. The application of linearly swept frequency measurements. *J. Acoust. Soc. America*, 84(2), 1988.
16 J.N. Mourjopoulos. On the variation and invertibility of room impulse response functions. *J. Sound and Vibration*, 102(2), 1985.
17 O. Kirkeby and P.A. Nelson. Digital filter design for inversion problems in sound reproduction. *J. Audio Eng Soc.*, 47(7/8), 1999.