Perceptual Evaluation of Numerical Auditory Scene Synthesis Using Loudspeaker Arrays

Ismael Nawfal, Joshua Atkins, Daniele Giacobello, and Stephen Nimick

Beats Electronics LLC, Culver City, CA, 90232, USA

Correspondence should be addressed to Ismael Nawfal (ismael@beatsbydre.com)

ABSTRACT

In this paper, we address the problem of subjectively evaluating the spatial quality of a given sound event, or *auditory scene*, produced using an array of loudspeakers and various spatial reproduction methods. In particular, we focus our attention on recently introduced numerical methods that are capable of addressing some of the limitations of analytical methods of sound field reproduction. The evaluation is performed through a novel listening test methodology derived from a common multiple stimuli test for which we provide the implementation details on its adaptation to spatial audio evaluation. The testing is done in both anechoic and non-anechoic conditions, and the results obtained are statistically significant and clearly outline some of the benefits and limitations of the presented methods.

1. INTRODUCTION

The increasing shift of surround sound reproduction in the home environment toward line arrays of speakers, mainly in high-end televisions and sound bars, raises the question of how to effectively reproduce and evaluate different spatial reproduction methods for these systems.

The literature on spatial audio rendering can be divided into four areas: attempts at accurate reproduction of a wave-field (e.g., wave-field synthesis (WFS) [1] or near-field compensated higher-order ambisonics (NFC-HOA) [2]); attempts at accurate binaural reproduction at a particular listening position (e.g., crosstalk cancellation [3] or loudspeaker binaural rendering (LBR) [4, 5]); attempts to reproduce the perceptual attributes of a sound field using heuristic approaches (e.g., vector-base amplitude panning (VBAP) [6]); and numerical approaches to reconstructing a sound-field (e.g., mode-matching for HOA [7] and the equivalent source method (ESM) [8]). The numerical techniques aim to generate a set of filters for each loudspeaker in an array by minimizing the ℓ_2-norm of the error between the reproduced field and a target response at a few listening positions. Frequently this is done in a narrow-band sense, focusing on a single frequency, and the generation of a broadband filter is left as a separate optimization problem. In our previous work [5], we introduced *numerical auditory scene synthesis* (NASS), a flexible numerical method that allows for broadband filter design and perceptually relevant error metrics encompassing the LBR, mode-matching HOA, and ESM methods. The work in [5] provided a first step towards analyzing the perceptual implications of various norms, regularization techniques, and target functions but used a very simple metric which compared the rendered transfer function at a simulated listener's ear drum to the desired transfer function.

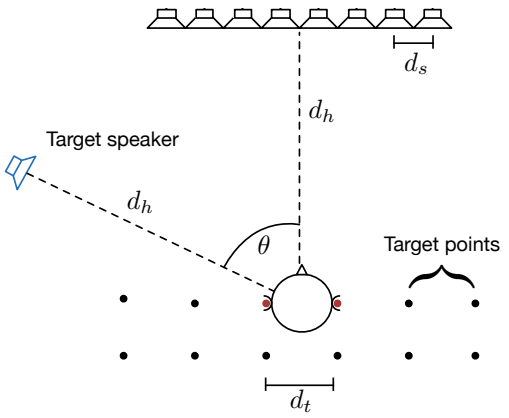


Fig. 1: Listening room setup for measurement of spatial audio rendering from a uniform linear array with a manikin. The values d_s and and d_h, represent, respectively, the distance between speakers and the distance from the center of the array to the target area. The red points show the target ear locations.

The literature on perceptual evaluation of the various spatial sound synthesis methods is somewhat sparse [9]. Historically, the standardization process on how to mea-

sure audio quality has focused mostly on distortions that affect the timbre, as caused by coding [10]. However, measuring the quality in spatial audio is related not only to the timbral characteristics, but also on localization and spatial extent [11]. How these multi-dimensional features are perceived affects the *character* of the auditory scene. One of the most notable efforts to provide a common methodology to evaluate spatial quality was done in [12]. However, the nature of the testing is limited to very small impairments where only expert listeners are involved in the evaluation. Furthermore, strict specifications are given on the experimental setup which make the testing expensive, both in terms of time and resources. Other types of standards, mainly [13], address the evaluation of spatial audio in the 5.1 format. However, it also restricts the location of the speakers to a particular configuration and is focused mainly on coding methods.

The most significant contribution in evaluating spatial quality in the recent past was proposed in [14, 15, 16, 17], where the quality evaluation of spatial transmission and reproduction using an artificial listener (QESTRAL) was introduced. The contribution of the QESTRAL methodology is twofold. Firstly, it introduces the idea of calculating the response at the ears of a manikin in any given position of a room to evaluate a simulated auditory scene with any loudspeaker configuration, thus greatly simplifying the experimental setup. Secondly, it provides a number of objective measures that map well to perception. As is done in other standardized objective quality evaluation techniques, e.g. [18, 19], the fusion of these measures and their mapping to a subjectively relevant score, or *calibration*, is done by analyzing a very large number of subjective experiments and their relation to different objective qualities. However, to date the QESTRAL method has only been validated on discrete 5.1 reproduction and has not been applied to evaluation of the previously listed methods for spatial audio synthesis.

This paper proposes a spatial perceptual evaluation methodology inspired by the QESTRAL listening tests to analyze the spatial quality of a selection of numerical auditory scene synthesis methods [5]. The method is based on simulating a binaural auditory scene using a single source from a given angle at the listener's ears, as shown in Figure 1. This allows for an inexpensive testing methodology through headphone playback based on the multiple stimuli with hidden reference and anchor (MUSHRA) paradigm [20].

The paper is organized as follows. In Section 2, we review various models of acoustical propagation used in many sound field synthesis techniques. The NASS method is presented in Section 3. In Section 4, we give details on the experimental setup for the perceptual testing and, in Section 5, we provide the results of the test and comparisons and relations to objective measures. Finally, in Section 6, we present our conclusions.

2. MODELS OF ACOUSTICAL PROPAGATION

Most methods of numerical sound field synthesis approach the problem in a similar way to their analytical counterparts by assuming a simplified model of acoustical propagation between each loudspeaker and each target point. The choice of acoustical model has a significant impact on the resulting system, both in terms of how well it matches the real-world listening environment (many models assume anechoic propagation, but are meant to be listened to in a reverberant room) and in terms of the intended use (wave-field methods are often used in systems for human listeners where head related propagation would be more fitting).

In this work, we generalize the acoustical propagation model to be any impulse response from source to receiver, whether it is measured or generated analytically. The generic impulse response is then written as $\mathbf{g} = [g_{\mathbf{r}}[n], \ldots, g_{\mathbf{r}}[n-N_g+1]]$ where N_g is the length of the measured impulse response. Analytical acoustical models are the common plane-wave, spherical-wave, and multipole models which, in free-field, are given by [21]

$$G(f) = Ae^{i(\mathbf{k}\cdot\mathbf{r}-\omega t)} \quad \overset{\mathscr{F}^{-1}}{\longrightarrow} \quad g(t) = A\delta\left(\frac{\mathbf{n}\cdot\mathbf{r}}{c} - t\right),$$

$$G(f) = \frac{Ae^{i(kr-\omega t)}}{r} \quad \overset{\mathscr{F}^{-1}}{\longrightarrow} \quad g(t) = \frac{A}{r}\delta\left(\frac{r}{c} - t\right),$$

$$G(f) = \sum_{l=0}^{L-1} \frac{A_l e^{i(k_l r-\omega t)}}{r} \quad \overset{\mathscr{F}^{-1}}{\longrightarrow} \quad g(t) = \sum_{l=0}^{L-1} \frac{A_l}{r}\delta\left(\frac{r}{c} - t\right),$$

respectively, where $k = \frac{\omega}{c}$ is the wavenumber, $\mathbf{k} = k\mathbf{n}$ is the wavenumber vector with unit vector $\mathbf{n}$ pointing in the direction of propogation, $\mathbf{r}'$ is the vector pointing from the origin to the source location, c is the speed of sound in the medium (approx. 343 m/s in air), $\omega = 2\pi f$ is the frequency (f is in Hertz), i is the imaginary number, $\mathbf{r}$ is the vector pointing from the origin to the evaluation point, $r = |\mathbf{r} - \mathbf{r}'|$ is the distance from source to evaluation point, $\delta(\cdot)$ is the Dirac delta function, and $\mathscr{F}^{-1}$ rep-

resents the inverse Fourier transform. These impulse responses are then sampled at discrete time steps and thus can have delays that are fractions of a given sample rate [22].

The system describing the acoustical field created from a set of S sources and measured at a set of M target points can be written as

$$\mathbf{y} = \mathbf{G}\mathbf{x}, \tag{1}$$

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1S} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{M1} & \cdots & \mathbf{G}_{MS} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_S \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_M \end{bmatrix},$$

where the acoustical propagation matrix $\mathbf{G}$ is composed of matrices $\mathbf{G}_{ms}$, each representing the $(N_g + N_x - 1) \times N_x$ dimensional acyclic convolution matrix of an individual impulse response $\mathbf{g}_{ms}$ and each loudspeaker signal $\mathbf{x}$ is $\mathbf{x}_s = [x_s[n], \ldots, x_s[n-(N_x-1)]]^T$.

3. FILTER OPTIMIZATION FRAMEWORK

We formalize the problem of determining a set of S filters $\mathbf{h}_s$, to apply to the source signal $\mathbf{x}$ to approximate a sound field as

$$\hat{\mathbf{y}} = \hat{\mathbf{G}}\mathbf{H}\mathbf{x}, \tag{2}$$

where $\hat{\mathbf{G}}$ is the measured acoustic transfer function used in (1), and $\mathbf{H}$ is a block diagonal matrix where each element is the convolution matrix of $\mathbf{h}_s$. The desired sound field can be represented as $\mathbf{y} = \mathbf{T}\mathbf{x}$, where $\mathbf{T}$ is the ideal response of the system shaped like $\mathbf{G}$ in (1), where each block represents the transfer function between the s-th source and m-th speaker. Thus, we can rewrite our problem as $\mathbf{T}\mathbf{x} \approx \hat{\mathbf{G}}\mathbf{H}\mathbf{x}$. Given the particular shape of the convolutional matrices, the problem becomes

$$\mathbf{G}\mathbf{h} - \mathbf{t} = \mathbf{e}. \tag{3}$$

We can now consider the optimization problem associated with finding a set of S filters $\mathbf{h}_s \in \mathbb{R}^{N_h}$ from a set of observed MS acoustic path models $\mathbf{g}_{ms} \in \mathbb{R}^{N_g}$ so that the reproduction error of the target function $\mathbf{t} = \left[\mathbf{t}_1^T, \ldots . \mathbf{t}_M^T\right]^T$, $\mathbf{t}_m \in \mathbb{R}^{N_t}$, where $N_t = N_g + N_h - 1$, is minimized.

The transfer function of a room can be exactly inverted for the case $N = 2M$, ($M = 1$ and $N_g = N_h + 1$) as was shown for the MINT method [23]. That is the case of $\mathbf{G}$ being a square matrix and thus the system having a unique solution, provided that $\mathbf{G}$ is full rank.

There is also an exact solution, when $NN_t \geq MN_h$ and the matrix has full row rank, $\mathrm{R}(\mathbf{G}) = MN_h$, a condition that can generally be assumed to be fulfilled for a convolutional matrix. In this case, the system is said to be *underdetermined* and it has infinitely many solutions, so we are seeking a particular solution, typically one that minimizes the ℓ_p-norm, defined as $\|\mathbf{x}\|_p = (\sum_{n=1}^{N} |x(n)|^p)^{\frac{1}{p}}$, of the solution vector. The optimization problem becomes:

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{h}\|_q \ \text{ s.t. } \ \mathbf{G}\mathbf{h} = \mathbf{t}. \tag{4}$$

However, since perfect multichannel inversion is hard to achieve when spatial robustness and possible perturbations of the measurement are considered [24], exploring the neighborhood of the minimum norm solution (4) and determining approximate solutions is of general interest. The condition $\mathbf{G}\mathbf{h} = \mathbf{t}$ can then be relaxed through the choice of an appropriate cost function. The optimization problem can then be generalized as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{W}(\mathbf{G}\mathbf{h} - \mathbf{t})\|_p \ \text{ s.t. } \ \begin{array}{l} \|\mathbf{Z}_i\mathbf{h}\|_{q_i} \leq \delta_i, \\ \forall i, i = 1 \ldots, I. \end{array} \tag{5}$$

where the matrices $\mathbf{Z}_i$ and $\mathbf{W}$ are added to represent linear projections, e.g., perceptual weighting, or transformations in a given domain. This method is called numerical auditory scene synthesis (NASS) and the implications of different projection matrices and norm choices are discussed in detail in [5]. Note that the minimization problem is the same in the case where $MN_h > NN_t$, when the system is *overdetermined*, and the condition $\mathbf{G}\mathbf{h} = \mathbf{t}$ cannot be fulfilled.

4. EXPERIMENTAL ANALYSIS

We focus our attention on the optimization problem assuming an ℓ_2-norm criterion on the unweighted cost function and impose a frequency domain constraint on the solution vector. The problem in (5) is then rewritten as

$$\hat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{G}\mathbf{h} - \mathbf{t}\|_2 \ \text{ s.t. } \|\mathbf{F}\mathbf{h}\|_q \leq \delta, \tag{6}$$

where $\mathbf{F}$ is a block diagonal matrix containing M DFT matrices and the m-th target vector is defined as

$$\mathbf{t}_m = [\underbrace{0, \ldots, 0}_{D}, t_m[0], \ldots, t_m[N_t - D - 1]],$$

where D represents the modeling delay that is used to ensure causality of the designed filter [24]. From (6), we derive six methods for generating a sound field. Three methods use a measured anechoic head-related impulse response (HRIR) as the acoustical propagation model and target, i.e. the loudspeaker binaural rendering (LBR) method presented in [4], and three methods use a free-field spherical wave model (SWM) for the acoustical propagation model and target similar to the ESM method [8]. We selected two underdetermined situations, one where the minimum ℓ_2-norm solution is selected and a second where the minimum ℓ_∞-norm solution is selected, along with one overdetermined situation where an ℓ_∞-norm constraint was used in the frequency domain:

- ℓ_2-norm LBR, underdetermined (LBR2_L2)
- ℓ_∞-norm LBR, underdetermined (LBR2_Li)
- ℓ_2-norm LBR, overdetermined, $\delta = 12$ dB (LBR12_Li)
- ℓ_2-norm SWM, underdetermined (WAVE2_L2)
- ℓ_∞-norm SWM, underdetermined (WAVE2_Li)
- ℓ_2-norm SWM, overdetermined, $\delta = 12$ dB (WAVE12_Li)

For all the methods, we chose $N_g = 256$ (5.3 ms), $N_h = 256$ (5.3 ms), $D = 100$ (2.1 ms), and $N_t = N_g + N_h - D - 1 = 411$ (8.6 ms) as defined in Section 3. The value of D was chosen to ensure causality of the designed filters [3, 4]. The values of N_g and N_h were chosen to encompass the direct portions of measured and target impulse responses respectively.

4.1. Experimental Setup

All methods were evaluated using a uniform linear array (ULA) consisting of eight speakers separated by 10 cm, corresponding to a spatial aliasing frequency of approximately 1700 Hz. Using this array, two simulated binaural listening tests were created: anechoic and non-anechoic. The general setup was identical for both the anechoic and non-anechoic cases and is illustrated in Figure 1 where the simulated centered listener position is shown along with the two and twelve target point locations used in the filter generation. The distance between the ears of the manikin head was approximately 15 cm and the distance between the center of the manikin head and the ULA was 2 m. The spatial quality was evaluated against a reference loudspeaker placed at 60° to the listener's left. This angle was chosen for two reasons: rendering of sources outside of the array's aperture become progressively more difficult the further out the source is and the 60° angle has interesting commercial applications requiring the widening of a sound stage.

The HRIRs used both in the LBR filter design and for anechoic simulations were measured in an anechoic chamber with a cutoff of approximately 40 Hz using a KEMAR manikin placed 3 m from a studio monitor with a flat frequency response from 50 Hz to 20 kHz. A total of 3600 measurements were taken at 0.1° increments and impulse responses were acquired using the logarithmic sine sweep method [25].

In contrast, the HRIRs used in the non-anechoic simulation were measured using a prototype 8 speaker array in a real room. The room used was 6.5 x 4.25 x 2.75 m with absorption and diffusion paneling on walls, a carpeted floor, and a broadband reverberation time (RT60) of 0.23 s. The room was designed meet the listening room requirements for acoustical propagation in [12] except for the noise floor requirement. However, since all subjective tests were carried out using binaural simulations on headphones and not with participants listening in the room this did not cause an issue. A target utilizing the same loudspeaker as in the ULA was placed at 60° to serve as the reference in the listening tests.

4.2. Objective Evaluation

Objective simulation results are shown in Figure 2. The graphs show (from left to right) the wave field at 500 Hz, the frequency response of the generated filters for each of the loudspeakers, and the response at the ears of the manikin head for each of the six methods referred to above. The LBR underdetermined cases tend to closely match the expected ear responses at the central listening position, while the spherical wave underdetermined methods, though generating the expected acoustical waveform, do not match the expected responses. This is due to the spherical wave models not incorporating frequency dependent cues contributed by the head shadowing. However, the underdetermined cases do not appear to be spatially robust, as the filters are optimized exclusively for the central listening position.

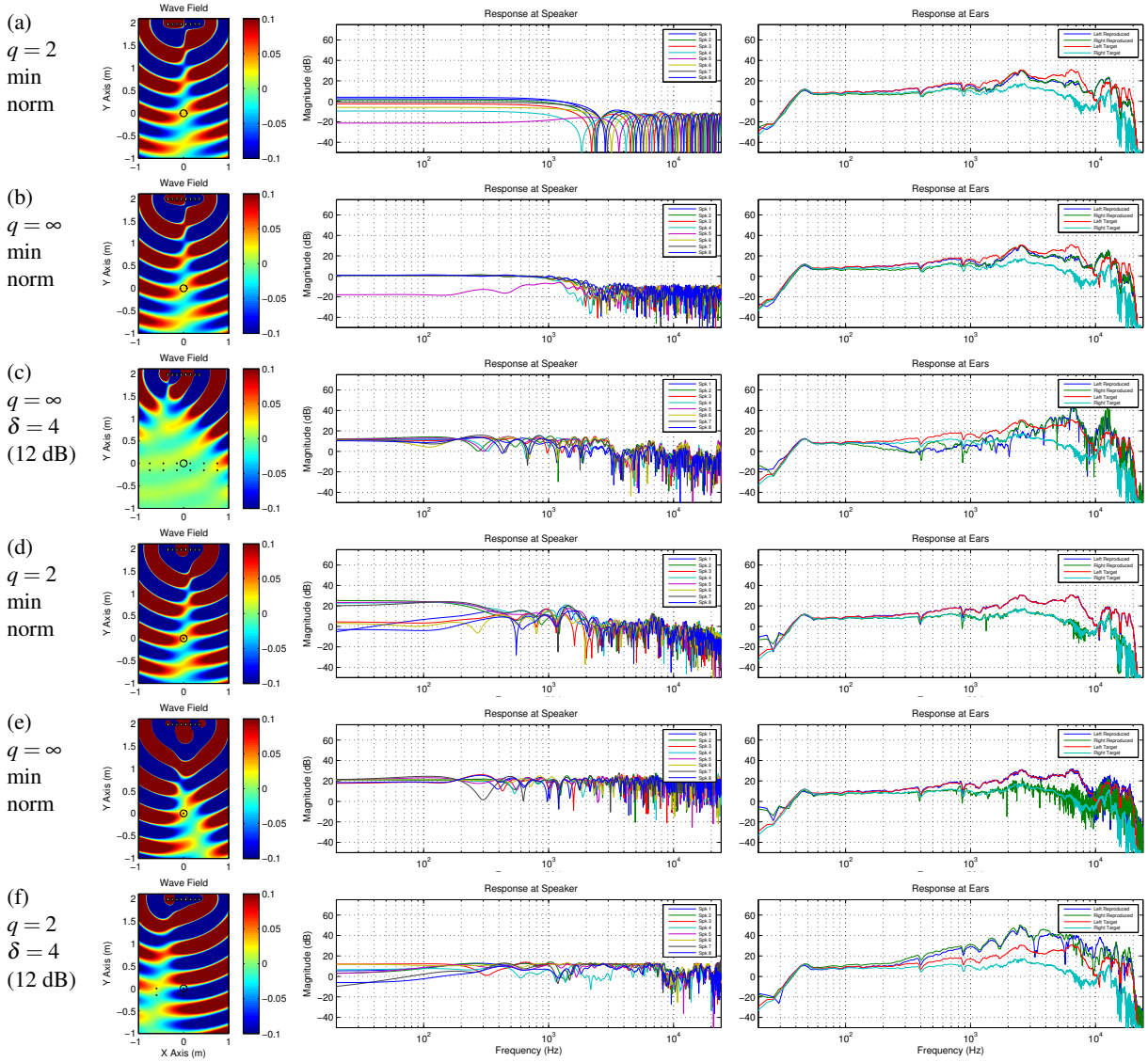


Fig. 2: Simulation results of a source rendered at 60° with an 8 loudspeaker array with $M = 2$ or $M = 12$ control points (black dots) for different values of q and δ. Subfigures (a)-(c) show results using a spherical wave acoustic propagation model, while (d-f) utilize measured HRTF responses. The left pane shows the wave field at 500 Hz; the middle and right pane show, respectively, the frequency response of the generated filters and the response at the ears of the manikin head (black circle). Subfigures (a), (b), (d), and (e) represent underdetermined systems ($M = 2$), while (c) and (f) represent overdetermined systems ($M = 12$).

In the overdetermined cases, the filters are optimized for a larger spatial region using 12 points. Using more target points increases the error between the resulting responses and desired responses at the ears. As noted in [5], a constraint was applied in order to obtain physically meaningful solutions considering the loudspeaker output. In particular, applying an ℓ_∞-norm constraint in the frequency domain allows for bounding the maximum value of the frequency response, as can be seen in Figure 2.

4.3. Subjective Evaluation

The standard metric for measuring the perceived quality of speech and audio signals is the mean opinion score (MOS) [26]. The measurement process consists of a pool of human listeners that compares the system under test

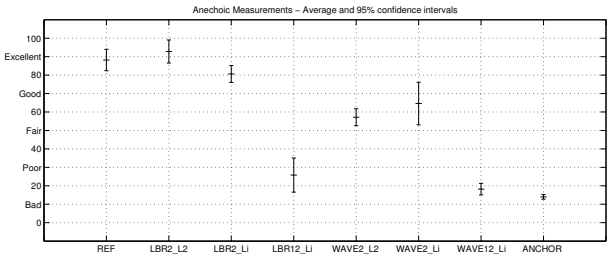
with a high quality fixed reference and ranks the degradation from "Inaudible" to "Very annoying" on a five-point scale. The MUSHRA test has similar scope, but the main advantage over the MOS methodology is that it requires fewer participants to obtain statistically significant results. Since all excerpts are presented at the same time, a paired t-test can be used for statistical analysis [20]. The participants are asked to rank the similarity of each file to a reference from a scale of 0-100.

Five diverse audio excerpts were evaluated: castanets, pink noise, music, male voice, and female voice. The audio excerpts used in the MUSHRA test were generated as perceived in a central listening location 2 m from the proposed ULA in an anechoic and non-anechoic room. Similar to the QESTRAL methodology, the listeners are not in these actual rooms, which allows for constant head placement and simple switching between algorithms. Thus, the simulations were conducted by measuring HRTFs associated with an 8-speaker ULA and target speaker in both types of environments using a KEMAR manikin. The appropriate HRTFs were then convolved with each speaker output to simulate the the actual acoustical propagation and head related transfer functions. Appropriate headphone compensation was used to remove the effects of the headphone on the binaural listening task. A headphone compensation filter was created by generating a smoothed inverse filter from 10 reseated KEMAR headphone measurements.

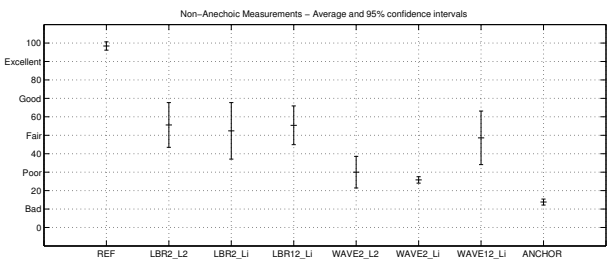
In order to remove the influence of loudness during perceptual evaluation, the loudness (loudness, k-weighted, relative to full scale, or LKFS) of all simulations was normalized based on [27]. This was achieved by calculating the average loudness of files processed with each of the six rendering techniques in both anechoic and non-anechoic simulations and comparing against their respective reference loudnesses. The difference between the average loudness and the reference loudness was used to normalize each set of files for playback.

The anchor was generated by low-pass filtering the original signal at 3.5 kHz, and a simple decorrelator was used to reduce the anchor's perceived directionality.

The test consisted of 13 listeners; 9 experts who had considerable musical and audio engineering training and 4 who were considered naive listeners. The participants were instructed to compare all files to a labeled reference and were informed that there was a hidden reference and anchor within the evaluation. The participants were intentionally not asked to evaluate specific audio qualities such as timbre and spatial focus; this was done to avoid misinterpretation of the experiment and to strictly focus the experiment on the general perceptual difference between references and test files.



(a) Anechoic



(b) Non-anechoic

Fig. 3: Anechoic (a) and non-anechoic (b) MUSHRA results for the spatial reproduction methods evaluated.

5. RESULTS

The results of the MUSHRA test in the anechoic scenario are shown in Figure 3.a. The underdetermined systems rank significantly higher than the overdetermined ones given that the target points coincide with the location of the dummy ears. Of the underdetermined cases, LBR cases scored highest when compared to all other methods with significant overlapping of the confidence region with that of the reference. This indicates the importance of spectral cues embedded in the LBR method, which are not included in the spherical wave propagation methods. The overdetermined cases were not ranked favorably for both the LBR and wave model methods due to the minimization effort targeting a much larger area (12 points). These results are particularly interesting given the fact that the single frequency wave fields plotted in Figure 2 look correct at the listening position for many of the methods.

The results of the MUSHRA test using a non-anechoic room simulation, shown in Figure 3.b, generally show lower scores, agreeing with the mismatch between the anechoic calculation and actual environment. In these cases, the overdetermined LBR and wave propagation methods scored more favorably than in the anechoic room simulation test. This suggests that optimizing over a larger space allows for relaxation of spatial constraints and reduction of adverse room effects.

Though beyond the scope of this work, the clear decay in score between anechoic and non-anechoic cases suggests that the characteristics of a listening room should be taken into account during the filter generation process in order to accurately render sources at arbitrary angles, as done in, e.g., [28, 29].

In Figure 4, the mean MUSHRA scores are plotted against the calculated Log Spectral Deviation (LSD) between the target and actual frequency-domain magnitude responses at a listener's ears for all six methods. The anechoic MOS and LSD data had a correlation coefficient of -0.78 which suggests that there is a reasonably strong relationship between MOS and LSD data derived from the anechoic room. This should be expected as lower LSD would imply a closer match to the target response at the ears which would presumably result in higher MOS scores. However, this is not true in the non-anechoic case where the correlation between MOS scores and LSD is not particularly significant. This motivates the search for more perceptually relevant cost functions and constraints for numerical auditory scene synthesis, many of which can be achieved through the existing framework via the existing projection matrices.

It should be noted that the MUSHRA scores obtained do not explicitly indicate whether the judged impairments of each technique were spatial or timbral in nature when compared to their respective reference. When informally evaluating the test clips it was clear that both timbral and spatial impairments were present to varying degrees in each technique.

6. CONCLUSION

In this paper, we compared a newly proposed numerical method, LBR, that takes into account the acoustical propagation from a loudspeaker array to a listener using an HRIR to the more standard spherical-wave model used in many spatial audio synthesis methods. Broadband filters

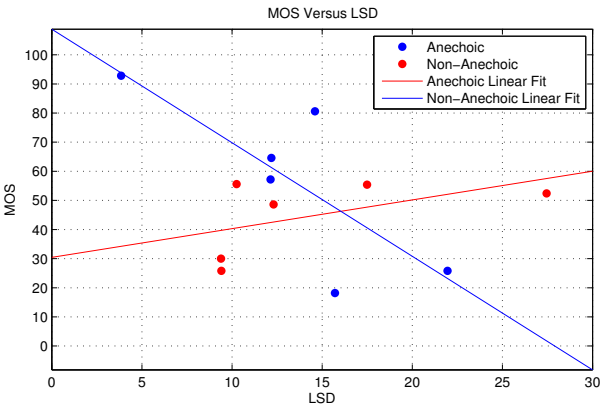


Fig. 4: MUSHRA scores versus LSD for anechoic and non-anechoic rooms. The anechoic room showed a correlation coefficient of -0.78 while non-anechoic was 0.53.

were designed in both cases using the NASS method for a variety of cost functions and target configurations. The comparison was done through subjective listening tests in both anechoic and non-anechoic environments. The results indicate that LBR methods perform better than spherical-wave simulations at a centered listening point given the stronger ability of reproducing the broadband spatial cues. However, the results clearly indicate a mismatch between anechoic algorithm design and deployment in realistic environments. Furthermore, the weak relation between perceptual scores and objective scores, such as LSD, in the non-anechoic case suggests that the mean-square optimality, or ℓ_2-norm, that generally drives the optimization algorithm is not always justified. Ultimately, this suggests that new optimization metrics that relate to perceptual quality measures should be used in the optimization problem and new criteria investigated.

7. REFERENCES

[1] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *J. ASA*, vol. 93, p. 2764, 1993.

[2] J. Daniel, "Spatial sound encoding including near field effect: introducing distance coding filters and a viable, new ambisonic format," in *Proc. 23rd AES Conf.*, 2003.

[3] D. B. Ward, "Joint least squares optimization for robust acoustic crosstalk cancellation," *IEEE Trans. SAP*, vol. 8, no. 2, pp. 211–215, 2000.

[4] I. Nawfal and J. Atkins, "Binaural reproduction over loudspeakers using a modified target response," to appear in *Proc. ICAD*, 2014.

[5] I. Nawfal, J. Atkins, and D. Giacobello, "A unified approach to numerical auditory scene synthesis using loudspeaker arrays," to appear in *Proc. EUSIPCO*, 2014.

[6] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. AES*, vol. 45, no. 6, pp. 456–466, 1997.

[7] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *J. AES*, vol. 53, no. 11, pp. 1004–1025, 2005.

[8] G. H. Koopmann, L. Song, and J. B. Fahnline, "A method for computing acoustic fields based on the principle of wave superposition," *J. ASA*, vol. 86, no. 6, pp. 2433–2438, 1989.

[9] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial sound with loudspeakers and its perception: a review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, 2013.

[10] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "PEAQ – the ITU standard for objective measurement of perceived audio quality," *J. AES*, vol. 48, no. 1/2, pp. 3–29, 2000.

[11] T. Lund, "Enhanced localization in 5.1 production," in *Proc. 109th AES Conv.*, 2000.

[12] *Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*, ITU-R Rec. BS.1116-1, 1997.

[13] *Multichannel stereophonic sound system with and without accompanying picture*, ITU-R Rec. BS.775-3, 2012.

[14] F. Rumsey, et al., "QESTRAL (part 1): quality evaluation of spatial transmission and reproduction using an artificial listener," in *Proc. 125th AES Conv.*, 2008.

[15] R. Conetta, et al., "QESTRAL (part 2): calibrating the QESTRAL model using listening test data," in *Proc. 125th AES Conv.*, 2008.

[16] P. Jackson, et al., "QESTRAL (part 3): system and metrics for spatial quality prediction," in *Proc. 125th AES Conv.*, 2008.

[17] M. Dewhirst, et al., "QESTRAL (part 4): test signals, combining metrics, and the prediction of overall spatial quality," in *Proc. 125th AES Conv.*, 2008.

[18] *Perceptual Objective Listening Quality Assessment (POLQA)*, ITU-T Rec. P.863, 2010.

[19] *Method for Objective Measurements of Perceived Audio Quality (PEAQ)*, ITU-R Rec. BS.1387-1, 2011.

[20] *Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)*, ITU-R Rec. BS.1534-1, 2003.

[21] E. G. Williams, *Fourier acoustics*, Academic Press, 1999.

[22] V. Välimäki and T. I. Laakso, "Principles of fractional delay filters," in *Proc. ICASSP*, vo. 6, pp. 3870–3873, 2000.

[23] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. ASSP*, vol. 36, no. 2, pp. 145–152, 1988.

[24] J.O. Jungmann, R. Mazur, M. Kallinger, T. Mei, and A. Mertins, "Combined acoustic mimo channel crosstalk cancellation and room impulse response reshaping," *IEEE Trans. ASLP*, vol. 20, no. 6, pp. 1829–1842, 2012.

[25] A. Farina, "Simultaneous measurement of impulse response and distortion with a swept-sine technique," in *Proc. 108th AES Conv.*, 2000.

[26] *Methods for subjective determination of transmission quality*, ITU-T Rec. P.800, 1996.

[27] *Algorithms to measure audio programme loudness and true-peak audio level*, ITU-R Rec. BS.1770-3, 2012.

[28] A. Canclini, et al., "A geometrical approach to room compensation for sound field rendering applications," to appear in *Proc. EUSIPCO*, 2014.

[29] P. Samarasinghe, et al., "Room reflections assisted spatial soundfield reproduction," to appear in *Proc. EUSIPCO*, 2014.