Audio Engineering Society

Convention Express Paper 167

Presented at the 155th Convention
2023 October 25–27, New York, USA

Excitation Stimuli For Simultaneous Deconvolution of Room Responses

Sunil Bharitkar[1], Ema Souza-Blanes[1], and Pascal Brunet[1]

[1]*DMS-Audio Lab, , Samsung Research America*

Correspondence should be addressed to Sunil Bharitkar (`s.bharitkar@samsung.com`)

ABSTRACT

This paper compares three state-of-the-art stimuli (multitone-pink, MLS, and log-sweep) to simultaneously deconvolve the impulse responses from several loudspeakers. A hyperparameter optimization algorithm constructs the stimulus, where the algorithm optimizes the stimulus parameters by minimizing a *time domain error* between the actual impulse responses and the simultaneously deconvolved responses over a training dataset. Objective results are presented for the various stimuli in a test data set that demonstrate the efficacy of each stimulus in the context of simultaneous deconvolution.

1 Introduction

Loudspeaker-room equalization begins with the acquisition of a loudspeaker-room impulse response, which entails recording the sound signal produced by a loudspeaker at a given position to the listener's location. The current approach involves extracting the response $h_{i,j}(n)$, obtained after energizing loudspeaker i with a stimulus and measuring at microphone position j ([5]-[18]). This process of deconvolution is repeated for each loudspeaker. Common stimuli employed for capturing room responses include pink noise which is commonly used for cinema calibration [19]; maximum length sequence (MLS) due to its well-understood mathematical properties [33]; and log-sweep due to its advantages over other stimuli [21].

However, a drawback becomes evident when dealing with a larger number of loudspeakers and positions, as the time needed to deconvolve responses from all loudspeakers becomes significant. Moreover, performing repeated measurements to enhance the signal-to-noise ratio (SNR) contributes to a prolonged calibration time. Recent strategies to address this limitation include the work by Majdak *et al.* [23] and Weinzierl *et al.* [24], which involve interleaving or partially overlapping sweep stimuli. Bharitkar [3] proposes a technique (illustrated in Fig. 1) for simultaneous deconvolution by exciting all loudspeakers *concurrently*. The log-sweep stimulus is optimized using Bayesian optimization, considering the *log spectral distortion metric* between the actual and estimated magnitude responses over different durations and circular shifts. This deconvolution method is validated in real-world scenarios [2], encompassing rooms with varying measured SNRs.

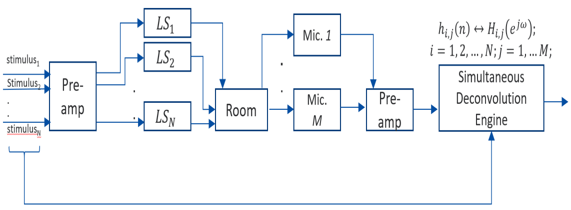

Fig. 1: Presented approach of simultaneous room response, $h_{i,j}(n)$, deconvolution for all loudspeakers ($LS_i; i = 1,2,...,N$) and microphone position j.

The key advancements in this paper are (i) incorporating impulse response error minimization in the Bayesian optimization enabling time-delay estimation and dereverberation, (ii) Bayesian optimization and comparisons between three popular excitation signals including the MLS (white), multitone (pink), and log-sweep for simultaneous deconvolution, (iii) objectively comparing the three stimuli using statistical analysis on the test set for generalization ability. Section II summarizes the properties of the three excitation stimuli used in this paper, the basic principles of simultaneous deconvolution and the dataset creation approach. Section III presents the Bayesian algorithm for optimizing the excitation stimuli parameters using the impulse responses for training and testing. Section IV presents objective results on the modeling performance with each stimulus over a test set, whereas Section V concludes/summarizes the paper.

2 Excitation Stimulus

2.1 Multitone (pink spectrum)

An input signal utilized for system identification involves a multitone waveform [29]-[32], characterized by specific amplitude, frequency, and a stochastic phase arrangement and is expressed as $u(t) = \sum_{k=-N/2+1}^{N/2-1} U_k e^{j\omega_k t}$. The phases $\angle U_k$ are random and uniformly distributed on $[0, 2\pi[$. This phase distribution ensures a signal with random values and an amplitude distribution that tends asymptotically to a Gaussian law when $N \to \infty$.

2.2 Maximum-length sequence

A Maximum-Length Sequence (MLS) [33] refers to a periodic signal characterized by two discrete levels, possessing a length of $P = 2^L - 1$, where L is an integer indicating the sequence length, and P represents its periodicity. The impulse response is extracted through correlation methods or the application of the Fast Hadamard Transform.

2.3 Logarithmic Sine-sweep

In the case of an exponential sweep, [34], assuming ω_1 and ω_2 being the start and end frequencies, with a total duration of T_{log} seconds, the logarithmic sweep signal $x(t)$ is

$$x(t) = \sin\left(\frac{\omega_1 T_{\log}}{\log\frac{\omega_2}{\omega_1}}\left(e^{\frac{t}{T_{\log}}\log\left(\frac{\omega_2}{\omega_1}\right)} - 1\right)\right) \tag{1}$$

whereas the discrete-time equivalent is $\mathbf{x}_1(n) = (x(n), x(n-1), x(n-2), \ldots, x(n-(P-1)))^T$, T represents the vector transpose and $P = T_{log}/T_s$ in samples[1],[2].

2.4 Simultaneous Deconvolution

The measurement (recording), assuming noiseless condition, is a linear convolution sum between the loudspeaker-room response $\mathbf{h}_i$ and the stimuli $\mathbf{x}_i(n)$,

$$\mathbf{y}(n) = \sum_{i=1}^{N=11} \mathbf{x}_i(n) \circledast \mathbf{h}_i \tag{2}$$

with

$$\begin{aligned} \mathbf{x}_1(n) &= [x(n), x(n-1), \ldots, x(n-P+1)]^T \\ \mathbf{x}_i(n) &= [x(<n-(i-1)M>_P), \\ &\ldots, x(<n-(i-1)M-1>_P), \\ &\ldots, x(<n-(i-1)M-P+1)>_P]^T; \\ &(i = 2, \ldots, 11) \end{aligned} \tag{3}$$

with $<m>_P = m$ modulo P, and $\mathbf{h}_i = [h_i(1), h_i(2), \ldots, h_i(K)]^T$ is a K-length impulse response. Bharitkar [3] presents a fast implementation involving computing the cross-spectrum between the measurement and excitation stimuli and the

[1] $T_s = 1/f_s = 1/48000$ (s), and f_s is the sampling frequency
[2] $T_{\text{stimuli}} = P_{\text{stimuli}}/48000$, where stimuli are either log-sweep, MLS, or multitone-pink

auto-spectrum of the excitation stimuli (appropriately circularly-shifted) to deconvolve room responses from loudspeakers excited simultaneously. Specifically,

$$\begin{aligned} S_{\mathbf{x}_j,\mathbf{x}_j}(e^{j\omega}) &= \mathscr{F}\{\mathbf{x}_j(n)\}\mathscr{F}\{\mathbf{x}_j(n)\}^* \\ S_{\mathbf{x}_j,\mathbf{y}}(e^{j\omega}) &= \mathscr{F}\{\rho_{(\mathbf{x}_j(n),\mathbf{y}(n))}\} = \mathscr{F}\{\mathbf{x}_j(n)\}\mathscr{F}\{\mathbf{y}(n)\}^* \\ \hat{H}_j(e^{j\omega)}) &= \frac{S_{\mathbf{x}_j,\mathbf{y}}(e^{j\omega})}{S_{\mathbf{x}_j,\mathbf{x}_j}(e^{j\omega})} \\ \hat{\mathbf{h}}_j &= \mathscr{F}^{-1}\{\hat{H}_j(e^{j\omega})\} \end{aligned} \tag{4}$$

2.5 Dataset Creation

The room impulse responses used in this paper are from MARDY [35][3] and MeshRiR [36][4] databases. The MARDY database has 72 loudspeaker-room responses obtained in a variable acoustics room with a Genelec 1029A loudspeaker. In contrast, MeshRiR has 14112 responses from a room with an array of 32 loudspeakers and a rectangular grid of 21×21 microphone positions[5]. Based on an augmented dataset (created by combining both databases), the number of 11-channel responses available for simulations is binomial $\binom{14112+72}{11} \approx 10^{38}$.

2.6 Bayesian Optimization

Bayesian optimization [39] is a global hyperparameter optimization technique, constrained on the bounds of the hyper-parameters, and is best suited for optimization with 20 or fewer hyper-parameters. The technique builds a surrogate function for the objective and quantifies the uncertainty in that surrogate using Gaussian process regression. Additionally, several parameters are required for initialization, including the type of acquisition function which guides the sampling for the optimal hyperparameters [40]. Recent advances can be found in [41]. In our optimization, we set 11 hyperparameters: (i) duration P, and (ii) right circular shifts M_i ($i = 1, \ldots, 10$).

3 Bayesian Optimization of Stimuli Parameters

For Bayesian optimization, a “training” dataset of size TR is created with 11-channel combinations of room impulse responses from the MARDY and MESHRiR databases. The responses are input to a Bayesian optimization process that optimizes the duration and inter-channel shifts by minimizing a metric $\bar{\psi}_{SD}^{\text{bayes}}$, where

$$\bar{\psi}_{SD}^{\text{bayes}} = \frac{1}{R}\sum_{k=1}^{R}\sqrt{\frac{1}{11}\sum_{j=1}^{11}\|\hat{\mathbf{h}}_j^{(k)} - \mathbf{h}_j^{(k)}\|_2^2} \tag{5}$$

is the root-mean-square error (RMSE) averaged over the training set of size $R = TR$. The box constraints for the search for the optimal duration and circular shifts during the Bayesian optimization process are $\{P_{low}, P_{up}\}$ samples and $\{M_{i,low}, M_{i,up}\}$ samples, respectively. Algorithm 1 is used for the optimization of the 11-channel stimuli hyperparameters, duration ($\hat{P}$) and circular shift ($\hat{M}_i; i = 1, \ldots, 10$), where the construction of the stimuli during each Bayesian optimization evaluation is given in (4).

Algorithm 1: Bayesian Optimization (BO) for Hyperparameter Search for Stimuli

Result: Stimuli(P^*, M_i^*),
$i = 1, \ldots, 10$; minimum : $\bar{\psi}_{SD}^{\text{bayes}}$

Initialize *bayesopt*: Construct base stimuli $\mathbf{x}_1(n)$ (4), Gaussian Process Active Set Size=*GPA*, Number of Seed Points=*NP*, Exploration Ratio=*ER*, box constraints $\{P_{low}, P_{up}\}$ samples and $\{M_{i,low}, M_{i,up}\}$ samples, TR, and true MARDY and MESHRiR responses $\mathbf{h}_j^{(k)}; j = 1, \ldots, 11; k = 1, 2, \ldots, TR$;
while *maxTime* $\leq T$ *seconds* **do**
  For each $\hat{P}$ and $\hat{M}_i$ candidate, construct 11-channel stimuli using (4);
  Compute the convolution sum (3) using true responses and excitation stimuli with candidate $\hat{P}$ and $\hat{M}_i$;
  Estimate the responses using (5);
  Update hyperparameters ($\hat{P}, \hat{M}_i$) using *bayesopt* to minimize $\bar{\psi}_{SD}^{\text{bayes}}$ using (6);
end
$T^*_{\text{stimuli}} = P^*/48000$ (seconds);

4 Results

For *each* stimuli, the box constraints during the optimization for the duration and circular shift were set em-

[3] https://www.commsp.ee.ic.ac.uk/ sap/
[4] https://github.com/sh01k/MeshRIR
[5] The number of responses is $32 \times 21 \times 21 = 14112$

pirically as $\{P_{low}, P_{up}\} = \{5, 30\} \times 48000$ (samples)[6], $\{M_{i,low}, M_{i,up}\} = \{4096, 131072\}; \forall i$ (samples). Additionally, $GPA^7 = 100$, $ER^8 = 0.5$ [38], $NP^9 = 10$, and $T = 259,200$ (s). The training set size is $TR = 500$, and the test set is of size $TS = 1000$, where each sample comprises 11 randomized responses per the dataset creation approach described in Sec. III.

4.1 Objective Results

As shown in Table 1, the shortest duration stimuli is log-sweep, which yields the smallest training set RMSE $\bar{\psi}_{SD}^{*,\text{log-sweep}} = 6.775 \times 10^{-6}$, whereas the RMSE for multitone-pink is $\bar{\psi}_{SD}^{*,\text{multitone-pink}} = 9.3592 \times 10^{-5}$, and the MLS $\bar{\psi}_{SD}^{*,\text{MLS}} = 8.943 \times 10^{-5}$. Also shown in Table 1 are the individual channel optimal right circular-shift value M_i (relative to channel 1) for each stimulus. The MLS and multitone-pink durations are similar. The advantage of short-duration stimuli includes a lower probability of insertion of impulsive noise during excitation. The present paper does not address immunity to steady-state noise (immunity which may be achieved using stimuli averaging). The generalization ability for each of the optimized stimuli is shown in Fig. 3 for the test set of size $TS = 1000$ where the y-axis is the averaged RMSE (computed using (6) with $R = TS$), with the 95% confidence interval of the mean, and expressed in dB. A log-sweep with random shift M_i and with $T_{\text{rand-sweep}} = T^*_{\text{log-sweep}} = 5.2379$ (s) result is also shown in Fig. 3 with the worst performance compared to the optimized stimuli. The best objective performance is achieved using the log-sweep stimuli with marginal differences in the 95% confidence interval ($\Delta_{\text{CI,log}} = 1.16$ dB, compared with $\Delta_{\text{CI,multi}} = 0.58$ dB, $\Delta_{\text{MLS,CI}} = 0.46$ dB, $\Delta_{\text{rand-sweep,CI}} = 0.43$ dB).

Table 1: Bayesian optimized parameters for stimuli

Param.	log-sweep	multitone	MLS
T^*_{stimuli}	5.2379 (s)	28 (s)	21.845 (s)
M^*_1	53886	78924	24308
M^*_2	85006	48686	85296
M^*_3	53256	64758	118423
M^*_4	89316	66214	46918
M^*_5	101774	83749	69150
M^*_6	78699	109905	130623
M^*_7	61029	6280	14266
M^*_8	92437	103992	10699
M^*_9	44056	55934	46022
M^*_{10}	18460	7271	5154

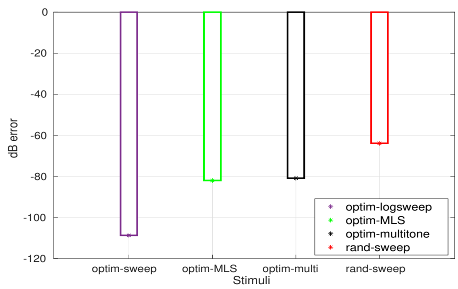

Fig. 2: Results from statistical analysis on $e_{\text{dB}}^{\text{test-set}}$.

5 Conclusions & Future Directions

This paper compares three widely-used stimuli for simultaneously exciting and deconvolving room responses. Each stimulus is optimized using Bayesian techniques for their duration, and circular shift amounts over a training dataset for an ITU 11-channel setup. The dataset is formed by augmenting the MARDY and MeshRiR databases creating a large corpus of 11-channel combinations. The objective performance (duration, MSE, and noise-robustness) demonstrates that log-sweep is the best candidate among the three stimuli over the test set using objective analysis. Future work will be done in the context of noise robustness and subjective preference of the stimuli. It may be possible to interpret the results using eigen-decomposition of various stimuli which will also be explored, along with comparisons with PSEQ [27] and [28].

[6]Based on footnote 3, the T_{stimuli} is box-constrained $\{5, 30\}$ (seconds)

[7]Fit Gaussian Process model to GPActiveSetSize or fewer points (using few points leads to faster GP model fitting, at the expense of possibly less accurate fitting)

[8]Parameter that balances between exploration and exploitation during global function optimization

[9]Number of initial evaluation points, specified as a positive integer, wherein bayesopt chooses these points randomly within the variable bounds

References

[1] S. Bharitkar, “Deconvolution of room impulse responses from simultaneous excitation of loud-

speakers," *Proc. 151st Audio Eng. Soc. Conv.,*, Online, Oct. 2021.

[2] R. Banka and S. Bharitkar, "Validation results of deconvolution of room impulse responses from simultaneous excitation of loudspeakers," *Proc. 153rd Audio Eng. Soc. Conv.,*, NY (USA), Oct. 2022.

[3] S. Bharitkar, "Bayesian Optimization for Simultaneous Deconvolution of Room Impulse Responses," *Proc. 36th IEEE Wkshp. Sig. Proc. Syst.*, Rennes (France), Nov. 2022.

[4] ITU-R BS. 2051-1, *Advanced sound system for programme production*, Int. Telecom. Union (ITU), 2018.

[5] S. Bharitkar and C. Kyriakakis, *Immersive Audio Signal Processing*, Springer-Verlag, June 2006.

[6] A. Carini, S. Cecchi, F. Piazza, I. Omiciuolo, G. Sicuranza,"Multiple position room response equalization in frequency domain," *IEEE Trans. Audio, Speech & Lang. Proc.*, 20(1), Jan. 2012, pp. 122–135.

[7] R. Mazur, F. Katzberg, A. Mertins,"Robust room equalization using sparse sound-field reconstruction," *2019 IEEE Int. Conf. Acoust., Speech & Sig. Proc.* (ICASSP 2019), Brighton (UK), April, 2019.

[8] M. Kolundžija, C. Faller, M. Vetterli,"Multichannel low-frequency room equalization using perceptually motivated constrained optimization," *2012 IEEE Int. Conf. Acoust., Speech & Sig. Proc.* (ICASSP 2012), Kyoto (Japan), 2012.

[9] M. Schneider and W. Kellermann, "Adaptive listening room equalization using a scalable filtering structure in the wave domain," *2012 IEEE Int. Conf. Acoust., Speech & Sig. Proc.* (ICASSP 2012), Kyoto (Japan), 2012.

[10] D. Menzies, P. Coleman, F. Fazi, "A room compensation method by modification of reverberant audio objects," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, Nov. 2020, pp. 239–252.

[11] J. Jungmann, R. Mazur, A. Mertins, "Joint time- and frequency-domain reshaping of room impulse responses," *2015 IEEE Int. Conf. Acoust., Speech & Sig. Proc.* (ICASSP 2015), Brisbane (Australia), 2015.

[12] S. Cecchi, A. Carini and S. Spors, "Room Response Equalization—A Review," *Applied Sciences*, 8(16), 2018.

[13] L. Fielder, "Practical Limits for Room Equalization," *Proc. 111th AES Conv.,*, NeY (USA), 2001.

[14] S. Bharitkar, C. Robinson, and A. Poulain, "Equalization of Spectral Dips Using Detection Thresholds," *Proc. 140th AES Conv.*, Paris (France), June 2016.

[15] C. Faller, "Modifying Audio Signals for Reproduction with Reduced Room Effect," *Proc. 147th AES Conv.*, NY (USA), Oct. 2019.

[16] D. Yellin and B. Friedlander, "Multichannel system identification and deconvolution: performance bounds," *IEEE Trans. Sig. Proc.* 47(5), 1999, pp. 1410–1414.

[17] T. Dobrowiecki, J. Schoukens, and P. Guillaume, "Optimized Excitation Signals for MIMO Frequency Response Function Measurements," *Proc. IEEE Instr. Meas. Tech. Conf.*,2005, pp. 1872-1877.

[18] F. Toole and S. E. Olive, "The Modification of Timbre by Resonances: Perception and Measurement," *J. Audio Eng. Soc.*, 36, 1988, pp. 122–142.

[19] ST 2095-1:2015, SMPTE Standard - Calibration Reference Wideband Digital Pink Noise Signal, Nov. 30, 2015.

[20] Y. Huang, J. Benesty. and J. Chen, "Identification of acoustic MIMO systems: Challenges and opportunities," *Sig. Proc.*, 86, 2006, pp. 1278–1295.

[21] G-B. Stan, J-J. Embrechts, and D. Archambeau, "Comparison of different impulse response measurement techniques," *J. Audio Eng. Soc.*, 50(4), April 2002, pp. 249–262.

[22] K. Prawda, S. J. Schlecht, and V. Valimaki, "Robust selection of clean swept-sine measurements in non-stationary noise," *J. Acoust. Soc. Amer.*, 151(3), March 2002, pp. 2117–2126.

[23] P. Majdak, P. Balazs, and B. Laback, “Multiple Exponential Sweep Method for Fast Measurement of Head-related Transfer Functions,” *J. Audio Eng. Soc.*, 55(7/8), July/Aug. 2007, pp. 623–637.

[24] S. Weinzierl, A. Giese, and A. Lindau, “Generalized multiple sweep measurement,” *Proc. 126th AES Conv.*, Munich (Germany), May 2009.

[25] C. Antweiler, A. Telle, and P. Vary, “NLMS-type system identification of MISO systems with shifted perfect sequences,” *Proc. IWAENC*, Seattle (USA), 2008.

[26] C. Antweiler, S. Kuhl, B. Sauert, and P. Vary, “System identification with perfect sequence excitation-efficient NLMS vs. inverse cyclic convolution,” *Proc. 11th ITG Conf. Speech Comm.*, Erlangen (Germany), 2014.

[27] C. Antweiler, A. Telle, P. Vary, and G. Enzner, “Perfect-sweep NLMS for time-variant acoustic system identification,” *Proc. Int. Conf. Acoust. Speech Sig. Proc.* (ICASSP), Kyoto (Japan), 2012.

[28] Y. Nakahara, Y. Iiyama, Y. Ikeda, and Y. Kaneda, “Shortest impulse response measurement signal that realizes constant normalized noise power in all frequency bands,” *J. Audio Eng. Soc.*, 70(1/2), Jan. 2022, pp. 24–35.

[29] I. F’ellejero, M. Zivanovic, I. Pmoabarren, A. Carlosena, “Application of multitone signals in room acoustics measurements," *Proc. IEEE Instr. & Meas. Tech. Conf.*, Budapest (Hungary), May 2001.

[30] J. Schoukens, R. Pintelon, “Identification of linear systems,” *Pergamon Press*, 1991.

[31] E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terekhov, “Multitone testing of sound system components–some results and conclusions, Pat 1: History and theory,” *J. Audio Eng. Soc.*, 49(11), Nov. 2001, pp. 1011–1048.

[32] A. Potchinkov, “Low-crest-factor multitone test signals for audio testing,” *J. Audio Eng. Soc.*, 50(9), Sept. 2002, pp. 681–694.

[33] J. Vanderkooy, “Aspects of MLS Measuring Systems,” *J. Audio Eng. Soc.*, 42(4), April 1994, pp. 503–516.

[34] A. Farina, “Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique,” *Proc. 108th Audio Eng. Soc. (AES) Conv.*, Paris (France), Feb. 2000.

[35] J. Wen, N. Gaubitch, E. Habets, T. Myatt, and P. Naylor, “Evaluation of speech dereverberation algorithms using the MARDY database,” *Proc. IWAENC*, Paris (France), Sept. 2006.

[36] S. Koyama, T. Nishida, K. Kimura, T. Abe, N. Ueno, and J. Brunnstrom, “MeshRIR: A dataset of room impulse responses on meshed grid points for evaluating sound field analysis and synthesis methods,” *Proc. IEEE-WASPAA*, Mohonk, NY (USA), 2021.

[37] L. van der Maaten, and G. Hinton, “Visualizing Data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, 2008, pp. 2579–2605.

[38] G. De Ath, R. M. Everson, A. Rahat, and J. E. Fieldsend, “Greed is Good: Exploration and Exploitation Trade-offs in Bayesian Optimisation,” *ACM Trans. Evol. Learn. Optim.*, 1(1), March 2021, pp. 1–22.

[39] J. Snoek, H. Larochelle, and R. Adams, “Practical Bayesian optimization of machine learning algorithms,” *Proc. Neural Inf. Proc. Syst. (NIPS)*, 2012.

[40] P. I. Frazier, “A tutorial on Bayesian optimization,” *arXiv:1807.02811v1*[stat.ML], July, 2018.

[41] J. R. Doppa, V. Aglietti, and J. Gardner, “Advances in Bayesian Optimization,” *Neural Inf. Proc. Syst. (NIPS) Wkshp.*, Dec. 2022.

[42] ITU-R BS. 1116, *Methods for the subjective assessment of small impairments in audio systems*, Int. Telecom. Union (ITU), 2015.

[43] ITU-R BS. 1770-4, *Algorithms to measure audio programme loudness and true-peak audio level* , Int. Telecom. Union (ITU), 2015.