MODAL ANALYSIS OF ROOM IMPULSE RESPONSES USING SUBBAND ESPRIT

Corey Kereliuk *
Reverberate.ca
St. John's, NL, Canada
info@reverberate.ca

Woody Herman
Eventide Inc.
Little Ferry, NJ
WHerman@eventide.com

Russell Wedelich
Eventide Inc.
Little Ferry, NJ
RWedelich@eventide.com

*Daniel J. Gillespie**
Newfangled Audio
New York, NY
DGillespie@eventide.com

ABSTRACT

This paper describes a modification of the ESPRIT algorithm which can be used to determine the parameters (frequency, decay time, initial magnitude and initial phase) of a modal reverberator that best match a provided room impulse response. By applying perceptual criteria we are able to match room impulse responses using a variable number of modes, with an emphasis on high quality for lower mode counts; this allows the synthesis algorithm to scale to different computational environments. A hybrid FIR/modal reverb architecture is also presented which allows for the efficient modeling of room impulse responses that contain sparse early reflections and dense late reverb. MUSHRA tests comparing the analysis/synthesis using various mode numbers for our algorithms, and for another state of the art algorithm, are included as well.

1. INTRODUCTION

Artificial reverberation is a now ubiquitous effect that is often used to add a sense of space and color to a live performance or recording. The acoustics of a reverberant space depend on several factors including a building's architecture, wall materials, furniture, and so on. These factors affect the intensity and directionality of echoes arriving at a listener over time. Artificial reverberation algorithms aim to model these echoes, either directly or indirectly, and often with different goals in mind as explained below.

Digital signal processing algorithms for artificial reverberation have a long history. A comprehensive examination of this history is given by the review article of Välimäki et al. [1]. A brief taxonomy of reverb algorithms includes:

- purely algorithmic and parametric approaches, e.g., Schroeder's allpass chains [2], feedback delay networks [3][4], sparse FIR filters [5], and modal filter banks [6] [7] [8],
- convolutional reverbs [9], and
- physical modelling [10].

The wide-variety of techniques for artificial reverberation is a testament to the importance of this effect. We may also conjecture that the development of different reverb algorithms has been led by different design goals. To illustrate, convolutional reverbs are capable of very accurate modelling[1] but are relatively inflexible. On the other hand, feedback delay networks are computationally efficient and easily modulated. The latter properties are important considerations when designing a reverb effect meant to act as an instrument in its own right [11].

An important concern of ours is the musicality/playability of the reverb, especially with respect to real-time manipulation of perceptually relevant qualities. At the same time, we desire a model that can accurately simulate real spaces[2]. These requirements led us to eschew the traditional convolution-based reverb in favor of a fully parametric approach. In particular, we have chosen to adopt a modal reverb architecture [6] because the mapping of modes to perceptually important parameters (room size, decay time), is relatively straightforward, and because the parameters of a modal filter bank can be stably modulated at audio-rate. Recent work has also demonstrated a variety of interesting techniques that can be used with modal filter banks for pitch processing, time-scaling, and distortion [12].

1.1. Previous work

Although modal architectures for reverb processing are relatively recent [6], similar techniques have been used in other contexts for quite some time. See for example: Laroche's model of heavily damped percussive sounds [13]; The source-filter piano model of Meillier et al [14]; Bank's instrument body model [8]; Paatero et al.'s modelling of loudspeaker responses [15]; and, Sirdey et al's modal analysis of impact sounds [16];

Within the realm of reverb effects several works address the estimation of modal parameters, including: the frequency zooming-ARMA model of Karjalainen et al. [7][17]; Abel et al.'s modal reverberator [6]; Maestre et al.'s pole optimization algorithm [18]; the Gabor ESPRIT model of Sirdey et al. [19]; Schoenle et al.'s model of room responses [20]; and Hashemgeloogerdi et al.'s work on subband Kautz-filter modelling [21].

1.2. Contributions

A particular problem with modal modeling of reverb is the high density of modes exemplary of real room responses. After a short duration, and above the Schroeder frequency, both the echo and modal density become so dense as to make estimation of explicit modes very difficult [22]. Even if we had access to these parameters, running a modal filter bank with more than a few thousand modes would unreasonably tax a typical CPU.

In order to confront the problem of modal estimation for very dense impulse responses we have chosen to use a high-resolution,

* For Eventide Inc.

[1] For a fixed source-listener positioning.

[2] In the same sense as a convolutional reverb.

parametric estimator: the ESPRIT algorithm [23] [24]. Due to its parametric nature, ESPRIT, does not suffer from the same resolution limitations encountered with Fourier transform-based estimators, e.g., [6]. Other works applying ESPRIT to estimate modal parameters include [25] [26] [19].

A difficulty with ESPRIT is that it becomes computationally intractable for very dense and very long responses, like those typically encountered for real rooms. For this reason, we have chosen to use a subband approach, which has several critical benefits as discussed in section 4.

In order to prune mode counts down to a realizable number for synthesis with a modal filter bank, our work presents an approach to reduce the model order using the K-means algorithm.

We also discuss a technique for managing early reflections, which are not always easy to model using a small number of modes.

1.3. Outline

The remainder of this paper is laid out as follows. Section 2 describes the synthesis model of the modal reverberator. Section 3 gives an overview and derivation of the ESPRIT algorithm. Section 4 gives an explanation of the subband modifications we've made to make ESPRIT tractable for such a large problem. Section 5 is a brief word on estimating the model order. Section 6 introduces our algorithm for fitting the initial magnitude and phase parameters of the modal reverberator. Section 7 shows how we reduce the number of modes while maintaining perceptual accuracy, while Section 8 describes an extension to handle early reflections. Section 9 describes 3 experiments we ran comparing this method to a ground truth, another algorithm, and with and without the special early reflection handling. Finally Section 10 shares conclusions and Section 11 contains references.

2. THE MODAL MODEL

A starting point for this work is the assumption that a measured room response, $h[n]$, can be perfectly modeled using a linear digital filter with a rational z-transform

$$H(z) = \frac{B(z)}{A(z)} = \frac{\sum_{k=0}^{N} b_k z^{-k}}{\sum_{k=0}^{M} a_k z^{-k}} \tag{1}$$

the poles of which correspond to roots of the polynomial $A(z)$. Using long division, followed by partial fraction expansion, we can re-write $H(z)$ as [27]:

$$H(z) = \underbrace{\sum_{k=0}^{N-M} B_k z^{-k}}_{H_{FIR}(z)} + \underbrace{\sum_{k=1}^{M} \frac{A_k}{1 - z_k z^{-1}}}_{H_{Modal}(z)} \tag{2}$$

which represents an FIR filter in parallel with a bank of 1-pole filters that define the resonant modes of the system. In the special case $N < M$, the FIR part disappears and $H(z) = H_{Modal}(z)$. We will assume this is the case for the time-being, and revisit the estimation of $H_{FIR}(z)$ in section 8.

Taking the inverse z-transform of $H(z) = H_{Modal}(z)$ gives

$$h[n] = \sum_{k=1}^{M} h_k[n] = \sum_{k=1}^{M} A_k z_k^n \tag{3}$$

assuming the impulse response is stable and causal. When the poles occur in complex conjugate pairs, the time-domain view of the modal filter bank represents an exponentially damped sinusoidal (EDS) model. The complex amplitudes $A_k = e^{\alpha_k + j\phi_k}$ define the initial magnitude and phase of each damped sinusoid $z_k^n = e^{(d_k + jw_k)n}$.

Given this model two goals remain: i) estimate the model order, M; ii) estimate the model parameters: initial magnitude, initial phase, frequency, and damping. The model order should be as small as possible, while still maintaining perceptual transparency of the impulse response.

3. ESPRIT

The ESPRIT algorithm can be used to find the frequency and damping parameters for the EDS model in equation (3). The seminal ESPRIT reference is [23], however it focuses on direction-of-arrival estimation for antenna arrays. A more recent reference that focuses specifically on audio signal processing is [24]. The ESPRIT algorithm is briefly described below.

First, we collect L samples of the impulse response $h[n]$ into a vector $\mathbf{h}$. We can then re-write the EDS model from (3) using vector matrix notation as follows

$$\mathbf{h} = \mathbf{E}\mathbf{a} \tag{4}$$

where $\mathbf{E}_{nk}$ and $\mathbf{a}_k$ correspond to z_k^n and A_k, respectively. Using the delay property:

$$z_k^{n+R} = z_k^R z_k^n \tag{5}$$

we can write the EDS model for the Hankel matrix $\mathbf{H}_{nk} = h[n+k]$ (consisting of delayed copies of $\mathbf{h}$) as

$$\mathbf{H} = \mathbf{E}\mathbf{A}\mathbf{E}^T \tag{6}$$

where $\mathbf{A}_{kk} = A_k$ is a diagonal matrix containing the complex amplitudes. The superscripts T and H indicate the matrix transpose and Hermitian transpose, respectively. The columns of $\mathbf{H}$ lie in the M-dimensional *signal space*, spanned by the modal vectors, i.e., the columns of $\mathbf{E}$. Although these are unknown, we can find another set of vectors that span the signal space via a singular value decomposition (SVD) of $\mathbf{H}$

$$\mathbf{H} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T \tag{7}$$

The column vectors of $\mathbf{U}$ are, in general, different from the signal vectors, however, they are related by an unknown linear transform $\mathbf{T}$ (a rotation and scaling)

$$\mathbf{E} = \mathbf{U}\mathbf{T} \tag{8}$$

The *rotational invariance* property of complex exponentials can now be invoked to determine the modal frequencies and dampings. Mathematically, the rotational invariance property states that

$$\mathbf{E}_{\uparrow} = \mathbf{E}_{\downarrow}\mathbf{D} \tag{9}$$

where $\mathbf{E}_{\uparrow}$ signifies deleting the first row of $\mathbf{E}$, $\mathbf{E}_{\downarrow}$ signifies deleting the last row of $\mathbf{E}$, and $\mathbf{D} = \text{diag}(z_0, z_1, \ldots, z_M)$. Substituting (8) into (9) and performing some algebra gives

$$(\mathbf{U}\mathbf{T})_{\uparrow} = (\mathbf{U}\mathbf{T})_{\downarrow}\mathbf{D} \tag{10}$$

$$\mathbf{U}_{\uparrow}\mathbf{T} = \mathbf{U}_{\downarrow}\mathbf{T}\mathbf{D} \tag{11}$$

$$\mathbf{U}_{\uparrow} = \mathbf{U}_{\downarrow}\underbrace{\mathbf{T}\mathbf{D}\mathbf{T}^{-1}}_{\mathbf{\Phi}} \tag{12}$$

$$\mathbf{\Phi} = (\mathbf{U}_{\downarrow}^H\mathbf{U}_{\downarrow})^{-1}\mathbf{U}_{\downarrow}^H\mathbf{U}_{\uparrow} \tag{13}$$

The matrix $\mathbf{\Phi}$ is computed using the Moore-Penrose pseudo inverse, since the matrix $\mathbf{U}$ is not typically square. The eigenvalues of $\mathbf{\Phi}$ are the complex modes $(z_1, z_2, \ldots, z_M)$, which can be recovered from an eigenvalue decomposition (EVD). Summarizing, the steps in the ESPRIT algorithm are:

1. Compute the signal space $\mathbf{U}$ [equation (7)]
2. Compute $\mathbf{\Phi}$ using the pseudo inverse [equation (13)]
3. Compute the complex modes from an EVD of $\mathbf{\Phi}$

4. SUBBAND PROCESSING

As alluded to previously, it is difficult to apply ESPRIT on long signals with high model orders because its complexity scales like $\mathcal{O}(LM(M + log(L)))$ [24].

One way to make ESPRIT tractable is to apply a divide and conquer approach. This can be done by passing the input through a filter bank to divide the input into a set of narrow subbands. There are four main benefits to this approach:

1. Since each subband has a narrow passband, we can safely assume that each subband contains a small number of significant modes. This in turn reduces the ESPRIT model order, M;
2. Using a suitable filter bank, we can downsample each subband without significant aliasing, which greatly reduces the amount of data, L, we need to consider when computing the SVD of the Hankel matrix;
3. Downsampling increases the distance between closely spaced modes, making them potentially easier to identify [7];
4. When using complex filters we can reduce the ESPRIT model order by a factor of 2 when analyzing real signals. During synthesis the complex conjugate modes can be restored to create a real impulse response.

Taken together, these aspects make it possible to apply ESPRIT to long IRs with potentially tens of thousands of modes. This approach was demonstrated in [19] using the Gabor transform and a similar idea was presented earlier by Laroche (using Prony's method instead of ESPRIT) [13].

We have experimented with three different filter bank architectures: the Gabor transform [19], the alias-free pyramidal filter bank described in [28], and the Audio FFT filter bank described in [29]. We currently use the Audio FFT filter bank in our analysis algorithm because it can be used to specify an arbitrary set of non-uniformly spaced subbands.

The r^{th} subband is produced by filtering the input with a causal N-tap FIR filter $g_r[n]$:

$$y_r[n] = h[n] * g_r[n] = \begin{cases} \sum_{k=1}^{M} \alpha_k \sum_{l=0}^{n} g_r[l] z_k^{n-l}, & \text{if } n < N-1 \\ \sum_{k=1}^{M} \hat{\alpha}_{kr} z_k^n, & \text{if } n \geq N-1 \end{cases} \tag{14}$$

where

$$\hat{\alpha}_{kr} = \alpha_k s_{kr} \tag{15}$$

$$s_{kr} = \sum_{l=0}^{N-1} g_r[l] z_k^{-l} \quad \text{(constant w.r.t. } n\text{)} \tag{16}$$

The first $N-1$ samples of the output $y_r[n]$ represent a start-up transient, which does not exhibit an EDS behavior. After the start-up transient dies out, however, each subband once again follows an EDS model, with the addition of a scaling factor s_{kr} that can be subsumed into the magnitude and phase for the current subband. For this reason, we ignore the first $N-1$ samples from each filter bank channel when applying ESPRIT on subbands. In our experience, this operation reduces the bias in the ESPRIT frequency and damping estimates. On the other hand, modes that have decay times comparable to the subband filter lengths cannot be accurately estimated.

For modes with center frequencies lying in the stopband of the r^{th} channel filter s_{kr}[3] should be negligibly small, allowing us to effectively ignore these modes in the current subband.

The Audio FFT filter bank's channel filters have been designed using the *window* method. It was demonstrated by [30] how the window method can be used to design perfect non-uniform reconstruction filter banks. We first choose R brickwall filters such that the sum of channel responses is unity

$$\sum_{r=1}^{R} G_r(e^{j\omega}) = 1 \tag{17}$$

where G_r is the frequency response of the r^{th} subband. This requirement is easily met by partitioning the frequency domain into a set of non-overlapping bands. Taking the inverse DTFT shows that

$$\sum_{r=1}^{R} G_r(e^{j\omega}) = 1 \iff \sum_{r=1}^{R} g_r[n] = \delta[n] \tag{18}$$

This set of filters is perfect reconstruction since we can recover the input signal $x[n]$ by adding together the subband responses, i.e.,

$$\sum_{r=1}^{R} y_r[n] = \sum_{r=1}^{R} x[n] * g_r[n] \tag{19}$$

$$= x[n] * \left(\sum_{r=1}^{R} g_r[n] \right) = x[n] * \delta[n] = x[n]. \tag{20}$$

However, due to the brickwall response of the channel filters each impulse response, g_r, is an IIR filter. Using the window method each channel IR is truncated via multiplication with a short window, creating an FIR filter. Using an N-tap window, $w[n]$, the r^{th} channel IR becomes $\hat{g}_r[n] = w[n] g_r[n]$. This set of filters is *still* a perfect reconstruction, if $w[0]$ is normalized to 1 since

$$\sum_{r=1}^{R} w[n] g_r[n] = w[n] \sum_{r=1}^{R} g_r[n] = w[n]\delta[n] = w[0]\delta[n] \tag{21}$$

Time-domain multiplication by $w[n]$ results in a convolution between the ideal channel filter and the window in the frequency-domain: $G_r(e^{j\omega}) * W(e^{j\omega})$. This results in a frequency-domain spreading of the filters, causing the filter responses to overlap in frequency. Figure 1 illustrates an example of this type of filter bank. The region marked as *partition* indicates the original boundaries of the ideal brickwall filter, and the region marked as *passband* shows the widened filter response due to the windowing. This particular filter bank was designed using a Chebychev window as suggested in [29].

[3] We recognize s_{kr} as the z-transform of the r^{th} subband filter evaluated at the k^{th} pole location.

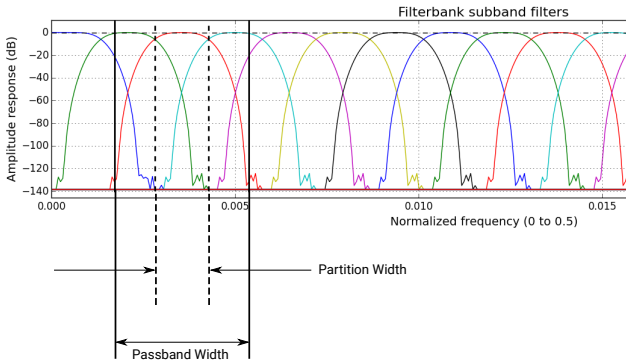


Figure 1: Filter bank design

When performing ESPRIT on subbands we can leverage the design of our filter bank in order to automatically prune out irrelevant modes. We first estimate how many modes are in each subband's *passband* as described in section 5 below. We then run ESPRIT using this model order. After ESPRIT returns we can safely discard any modes with center frequencies outside of the *partition*. We can do this because the partition perfectly divides the frequency spectrum into non-overlapping bands. Modes that do not lie in the current partition must belong to a neighboring partition (and therefore they should be estimated in the subband they lie closest to).

5. ORDER ESTIMATION

An inherent difficulty with parametric estimators lies in the specification of the model order—in our case the number of modes to estimate in each subband. There exist a few techniques that attempt to automatically estimate the model order based on information theoretic criteria, namely [31] and [32]. We have implemented these techniques, but found they did not perform particularly well for high model orders, e.g., more than 20 or so modes. Therefore, we have resorted to a simple model order selection algorithm based on peak picking from the discrete Fourier spectrum. We multiply the number of peaks detected by a *relaxation factor* greater than or equal to 1, recognizing the fact that some modes may not lead to a distinct peak in the sampled spectrum, or may be replicated (e.g., in the cases of two-stage and non-exponential decay). In practice, overestimating the model order does not usually pose a problem, because modes selected from the *noise subspace* generally have very small magnitudes.

6. MAGNITUDE AND PHASE ESTIMATION

After the modes z_k^n in each subband have been estimated using ESPRIT we must estimate the the complex amplitudes A_k. This can be done by minimizing the approximation error

$$\arg\min_{\mathbf{a}} ||\mathbf{h} - \mathbf{E}\mathbf{a}||_2^2 \tag{22}$$

A closed form solution to equation (22) is

$$\mathbf{a} = (\mathbf{E}^H\mathbf{E})^{-1}\mathbf{E}^H\mathbf{h} \tag{23}$$

However, this requires the inversion of a matrix with M^2 entries, which becomes very slow once the number of modes M exceeds a few thousand or so. We have experimented with conjugate gradient decent (which does not require a matrix inversion) to iteratively solve equation (22). This works well, but is still fairly slow once the number of modes exceeds several thousand.

Owing to Parseval's theorem, equation (22) can also be tackled in the frequency domain:

$$\arg\min_{\mathbf{a}} ||\mathbf{h} - \mathbf{E}\mathbf{a}||_2^2 = \arg\min_{\mathbf{a}} ||\check{\mathbf{h}} - \check{\mathbf{E}}\mathbf{a}||_2^2 \tag{24}$$

where $\check{\mathbf{h}}$ and $\check{\mathbf{E}}$ are the discrete Fourier transforms of $\mathbf{h}$ and the columns of $\mathbf{E}$, respectively. Note that each column of $\check{\mathbf{E}}$ can be computed analytically using the geometric series

$$\check{\mathbf{E}}_k[l] = \sum_{n=0}^{N-1} z_k^n e^{-j2\pi nl/N} \tag{25}$$

$$= \frac{1 - z_k^N}{1 - z_k e^{-j2\pi nl/N}} \tag{26}$$

In order to speed up the magnitude and phase estimation we once again resort to a divide and conquer approach. In particular, given a spectral filter $\mathbf{F}_k$ we can estimate the complex amplitudes of a subset of modes

$$\arg\min_{\mathbf{a_i}, i\in I_k} ||\mathbf{F}_k\check{\mathbf{h}} - \mathbf{F}_k\check{\mathbf{E}}\mathbf{a}||_2^2 \tag{27}$$

Modes that have minimal overlap with the filter $\mathbf{F}_k$ can be effectively ignored by removing columns from $\check{\mathbf{E}}$. Furthermore, we only need to minimize the norm in equation (27) over frequencies that fall in the passband of $\mathbf{F}_k$.

Using the DTFT we can calculate the 3dB bandwidth of the m^{th} mode to be

$$b_m = \arccos(2 - 0.5 * (e^{d_m} + e^{-d_m}))N/(2\pi) \tag{28}$$

where d_m is the damping factor and N is the DFT length. For the k^{th} subband we estimate the magnitude and phase of any modes for which the range $[\omega_m - b_m/2, \omega_m + b_m/2]$ intersects with the passband of the k^{th} spectral filter.

This procedure is applied repeatedly using a set of spectral filters $\{\mathbf{F}_k\}$ designed to completely cover the audible spectrum. This algorithm is much faster than any of the above techniques, and can be performed in parallel on architectures with multiple cores.

7. MODEL COMPRESSION

As mentioned in the introduction, in order to limit the CPU usage of a real-time modal reverberator we must restrict the total number of modes to no more than a few thousand. Subband ESPRIT routinely estimates upwards of 5000-10000 modes for real and dense IRs, meaning we require a strategy to reduce the overall number of modes used, ideally without sacrificing sound quality.

Luckily, it is possible to heavily compress our model by taking advantage of limitations in the human auditory perception system. In particular, it has been found that dramatically lower modal densities (compared to physically reality) can be used to generate perceptually accurate late reverberation. Therefore, we have developed a number of ad-hoc strategies to reduce the size of our modal filter bank

Following [18] we first partition the frequency spectrum into uniform bands on a Bark scale. We then divide our fixed modal

budget evenly across these bands. If some bands have fewer modes than they were allocated, the extra modes are reallocated among the remaining bands until no modes remain.

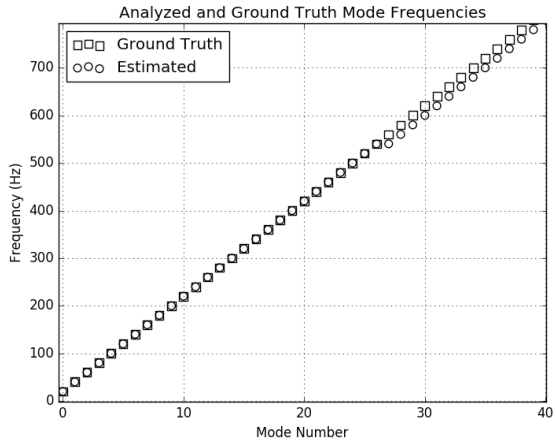
After the allocation step we have 2 numbers for each band i) M_a: the actual number of modes in each band (estimated using ESPRIT); and ii) M_d the desired number of modes in each band. While we could simply prune the extra modes $M_a - M_d$, this would change the distribution of modal frequencies in each band. Instead, we use the K-means algorithm to find a new set of M_d modes whose average distance from the estimated modes is minimized. An advantage of K-means is that is has the ability to 'average' the contributions of several modes by picking a new modal location that represents the center-of-gravity in a local neighborhood.

Empirically, we have found that the decay time estimates from ESPRIT exhibit a high degree of variance for real impulse responses. This in turn has a negative affect on the results of the K-means algorithm for small values of K (i.e., heavy model compression). In order to counteract this effect we smooth the decay time estimates from ESPRIT prior to running K-means. First, we apply a median filter to the decay times (after sorting them by frequency), which helps to eliminate outliers. Our median filter window starts with a length of 1 (at the boundaries) and grows until it reaches its maximum length (which is an algorithmic parameter in the range of 10 to several 100 modes). We have also experimented with weighted median filtering but no real benefit was noted. The median filtered decay times are then smoothed using a FIR lowpass filter to reduce the variance between nearby frequencies. It has been found that these three aspects: i) median filtering; ii) decay time smoothing; and, iii) K-means clustering are crucial for synthesizing *perceptually* good sounding IRs using a very small number of modes.

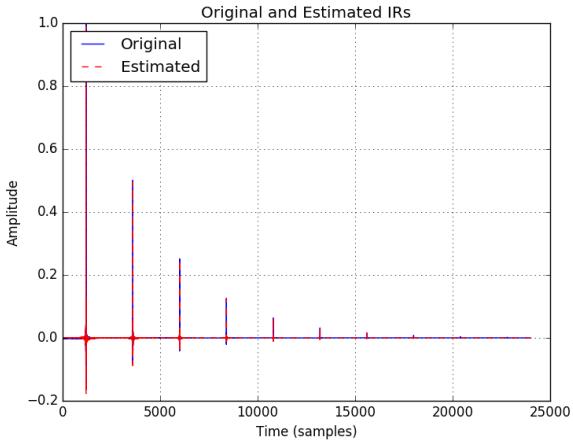
Once the model size is reduced the magnitude and phase of each mode (as discussed in section 6) must be re-estimated. In actuality, we always run the magnitude and phase estimation last, and hence only once.

We have applied a few additional ad-hoc strategies that should be noted. Immediately after running ESPRIT on each subband:

1. we discard any modes with a very low amplitude (estimated using least squares)
2. we discard any underdamped modes (which occur very rarely, and are unstable)



(a) Zoomed-in view of found modes.



(b) Synthesized and Estimated Irs.

Figure 2: Generated and found modes of a modally generated IR with 1000 modes.

8. HANDLING EARLY REFLECTIONS

Recall that our factorization of the rational transfer function in equation (2) included a parallel FIR path, $H_{FIR}(z)$. We can think $H_{FIR}(z)$ as modelling the early reflection portion of the reverb response. Fixing $H_{Modal}(z)$, the least squares solution for the FIR filter is $h_{FIR}[n] = h[n] - h_{Modal}[n]$ for $n \in [0, N)$. It is also possible to estimate the modal response from a delayed copy of the measured IR, i.e., $h[n - N_d]$. In this case

$$h_{FIR}[n] = \begin{cases} h[n] & \text{for} \quad n \in [0, N_d - 1] \\ h[n] - h_{Modal}[n] & \text{for} \quad n \in [N_d, N) \end{cases} \tag{29}$$

This later approach allows us to control the overlap between the responses which can lead to improved numerical performance as discussed in [33].

Before we can estimate the FIR part, however, we require some way to determine the tap-length of the FIR filter, N. In some cases, we have found that excluding the FIR part completely is a viable option, in which case we take $N = 0$. However, when an impulse response has prominent early reflections the modal synthesis algorithm may require an unreasonably large number of modes, M, to produce a good reconstruction on its own. We believe these unreasonably high mode counts originate from the EDS model's inability to efficiently model time sparsity[4]. A significant number of modes is required to build up the constructive/destructive interference pattern needed to model the sparsity between distinct echoes. In these situations we have implemented the early reflections using an FIR filter whose length, N, is estimated using Abel and Huang's echo density estimator [34].

[4]Indeed, the density of a Dirac comb in the time-domain is inversely proportional to its density in the frequency-domain.

9. EXPERIMENTS AND RESULTS

To validate the methods presented in this paper we conducted experiments with synthetic impulse responses having known modes, and performed several MUSHRA style listening tests. The recordings used in all of the listening tests are available online[5].

9.1. Synthetic Modal Impulse Response

In order to verify that the subband ESPIRIT analysis algorithm correctly identifies the true modes of an impulse response, we ran it on synthetically generated modal impulse responses with known sets of modes. The synthetic impulse responses were generated by adjusting the distribution of modes over frequency, the number of modes, and the decay times and magnitudes of the modes.

Figure 2 shows a plot of the known mode frequencies, obtained by spacing 1000 modes with a decay time of 0.5 seconds and magnitude of 1.0 linearly across the frequency spectrum up to 20kHz, as well as the modal frequencies detected by our subband ESPRIT analysis. Figure 2 aslo shows a plot of the original impulse response, and the one generated with modes found by subband ESPRIT. We can make the following observations: subband ESPRIT does indeed find the true modes of the impulse response, and it also finds a variety of spurious, or non-existent, modes. Note that even though the algorithm has added an extra mode at mode number 26, the subsequent mode frequencies are still correct. The addition of these spurious modes is, in part, due to the purposeful over-estimation of mode counts in the algorithm as previously discussed.

We can calculate the error between the known and estimated modal parameters by pairing the known and detected modes that are closest in frequency.

$$l_k = \arg\min_j(||f[k] - f_{\text{est}}[j]||_2) \tag{30}$$

$$e_f[k] = f[k] - f_{\text{est}}[l_k] \tag{31}$$

$$e_d[k] = d[k] - d_{\text{est}}[l_k] \tag{32}$$

Where l_k is the index of the detected mode that is closest to the k^{th} known mode in frequency. The mean and standard deviation of the error in the decay time estimates, e_d, are 0.000858 and 0.008301 *seconds* respectively. Similarly, we can calculate the mean and standard deviation of the errors between the known and found mode frequencies, e_f. These are $0.002329Hz$ and $0.015249Hz$, respectively. As a result of the close match between the modal parameters, the impulse response synthesized using the found modes is nearly identical to the one synthesized using the known modes. Comparing the two IRs, we find that the Mean Squared Error (MSE) between the two is -120.8147dB.

9.2. Monophonic Real Room Impulse Response

In order to compare our system with another state-of-the-art method, we chose to process the same impulse response presented in Maestre et al. [18] using mode counts of 400, 800, and 1800. In an effort to make the comparison fair, we did not include an FIR model of the early reflections in our model. We used the web-MUSHRA software [35] to administer a standard listening test in which users were asked to rate the quality of the modeled IRs with respect to a reference. A total of 12 users participated and there was no time-limit for the task. Figure 3 shows a box plot of the collected data, from which we may draw several conclusions. Firstly, the two algorithms perform quite similarly to one another. At low mode counts, our model was ranked slightly higher on average, whereas at high mode counts, Maestre et al.'s model was ranked higher. In our view, both models seem to impart very subtle artifacts to the impulse responses, however, test participants seemed to judge these artifacts differently depending on the model order.

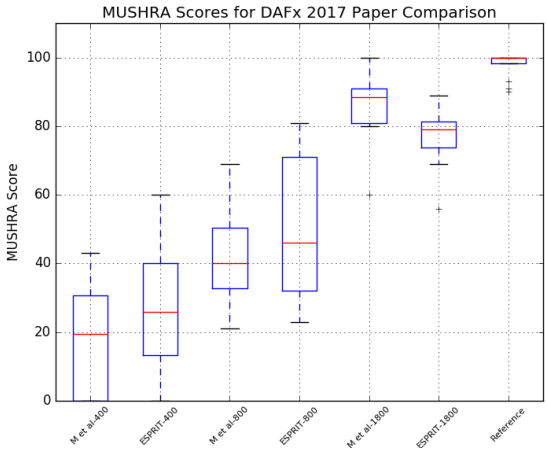


Figure 3: Listening test results I

Because the artifacts present in the synthesized impulse responses might manifest themselves differently when convolved with a source, we chose to perform a second test comparing our results with the results of Maestre et al. Using the same mode counts as before, listeners were asked to compare impulse responses which had been convolved with a source. The results of Maestre et al., and the dry source material, were obtained from their supplemental website[6]. Figure 4 shows the results of this test based on 9 users. Comparing Figures 3 and 4, two important observations stand out. Firstly, the scores of each individual response are, on average, higher than in the previous test and second, while our model was rated higher for a mode count of 800 in the previous test, the models of Maestre et al. were rated higher in this test.

9.3. Early Reflection Improvement with Parallel Synthesis Model

Figure 5 shows MUSHRA listening test results where 12 expert listeners rated the quality of differing subband ESPIRIT syntheses (and a hidden reference) with respect to a reference IR. The syntheses vary by model type, either pure modal or the FIR+modal model from Section 8, and the number of modes. In this experiment the reference is an IR from the Hall algorithm on a Lexicon PCM 90 digital reverb unit. This particular IR has a rather long early reflection field measuring 482 ms, measured using the Abel and Huang echo density estimator from [34]. The RT60 of this IR was also comparably long, measuring around 3 secs.

Listeners overwhelmingly favored the parallel FIR+modal model over the pure modal model. This trend held at very high

[5]http://dgillespie.github.io/Corey/

[6]https://ccrma.stanford.edu/ esteban/modrev/dafx2017/

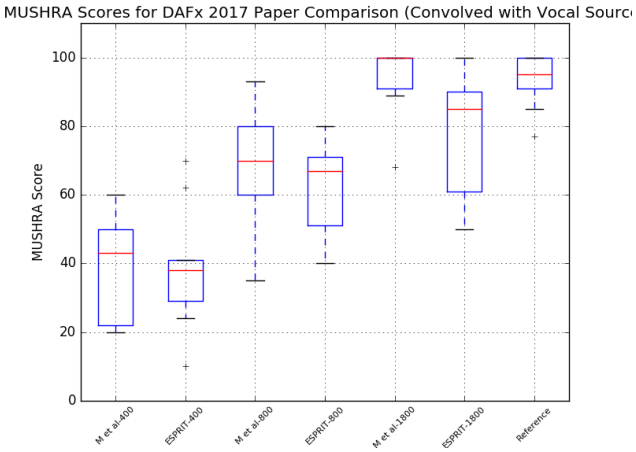


Figure 4: Listening test results II

mode counts for the pure modal model (12000). Even in this case, the parallel FIR+modal model using only 1500 modes rates significantly more similar to the reference. 12000 modes is taxing even on modern CPUs, while 1500 modes plus a fast convolution remains reasonably attainable.

We conclude that our hypothesis from Section 8 holds: it's difficult to guarantee accurate synthesis of significant early reflections in any efficient manner using a pure modal approach. Given that the early field is important in the perception and accurate synthesis of any given IR, the parallel model described in Section 8 can alleviate this particular issue. However, the parallel model is not without its drawbacks. The parallel model presents its own challenges for realtime audio effects like morphing, decay scaling, and size scaling because now the two parallel synthesis models must be parameterized in two differing domains and modulated in tandem to achieve perceptually pleasing and relevant results.

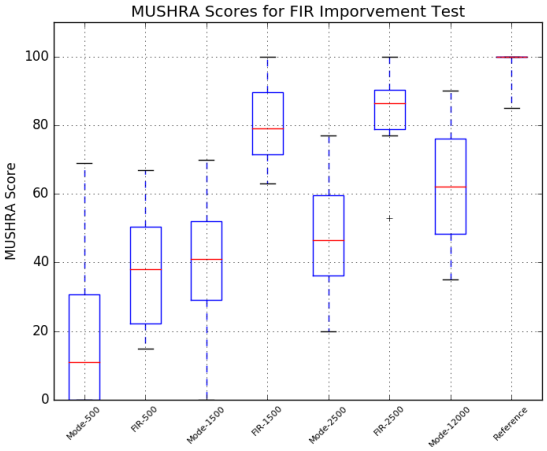


Figure 5: Listening test results III

10. CONCLUSION AND FUTURE WORK

In this paper we presented an end-to-end system for the modal analysis of real room impulse responses. Using the high-resolution ESPRIT estimator, we were are able to very accurately identify the frequency and damping parameters of impulse responses. Furthermore, we presented a number of strategies to i) make ESPRIT tractable on real-recordings; and, ii) yield models that can operate with fixed modal budgets. While our use of a subbband approach is not new, we have described several important considerations for practitioners of this method. This includes: trimming of the start-up transient, our approach to filter bank design, and our strategy for handling out-of-band modes. In order to reduce mode counts in the final model we presented a novel model compression algorithm based around K-means.

As mentioned previously, one interesting result of our listening tests was that, when convolved with a source, the results of Maestre et al. performed better than the method presented here. We have shown that the subband ESPRIT analysis method will find the correct modes of a system, given the correct model order, so it is likely that this error is introduced in our method of pruning excess modes. Because K-means is a clustering algortihm based on averages, the resulting set of modes after pruning may no longer be modes that were present in the original signal, but rather a new set of modes representing the average of several modes. Future work will surely focus on finding the best possible pruning method for reducing mode counts. This could include exploration of psychoacoustic-based methods, such as in [6]. In addition, it is worth noting that the rating of the generated impulse responses increased when they were convolved with a source. Presumably this is because some of the artifacts are masked. It would be of great value to know what artifacts are masked more heavily and vice versa. One could see the advantage in tuning the algorithm to be more accepting of artifacts that are more easily masked when used with source material.

11. REFERENCES

[1] Vesa Valimaki, Julian D Parker, Lauri Savioja, Julius O Smith, and Jonathan S Abel, "Fifty years of artificial reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1421–1448, 2012.

[2] Manfred R Schroeder, "Natural sounding artificial reverberation," *Journal of the Audio Engineering Society*, vol. 10, no. 3, pp. 219–223, 1962.

[3] John Stautner and Miller Puckette, "Designing multi-channel reverberators," *Computer Music Journal*, vol. 6, no. 1, pp. 52–65, 1982.

[4] Jean-Marc Jot and Antoine Chaigne, "Digital delay networks for designing artificial reverberators," in *Audio Engineering Society Convention 90*. Audio Engineering Society, 1991.

[5] Bo Holm-Rasmussena, Heidi-Maria Lehtonenb, and Vesa Välimäkib, "A new reverberator based on variable sparsity convolution," *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, vol. 5, no. 6, pp. 7–8, 2013.

[6] Jonathan S Abel, Sean Coffin, and Kyle Spratt, "A modal architecture for artificial reverberation with application to room acoustics modeling," in *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[7] Matti Karjalainen, Paulo AA Esquef, Poju Antsalo, Aki Mäkivirta, and Vesa Välimäki, "Frequency-zooming arma modeling of resonant and reverberant systems," *Journal of the Audio Engineering Society*, vol. 50, no. 12, pp. 1012–1029, 2002.

[8] Balázs Bank, "Direct design of parallel second-order filters for instrument body modeling.," in *ICMC*, 2007.

[9] William G Gardner, "Efficient convolution without input/output delay," in *Audio Engineering Society Convention 97*. Audio Engineering Society, 1994.

[10] Craig J Webb and Stefan Bilbao, "Virtual room acoustics: A comparison of techniques for computing 3d-fdtd schemes using cuda," in *Audio Engineering Society Convention 130*. Audio Engineering Society, 2011.

[11] Tom Erbe, *Building the Erbe-Verb: Extending the Feedback Delay Network Reverb for Modular Synthesizer Use*, Ann Arbor, MI: Michigan Publishing, University of Michigan Library, 2015.

[12] Jonathan S Abel and Kurt James Werner, "Distortion and pitch processing using a modal reverberator architecture," in *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx-15)*, 2015.

[13] Jean Laroche, "A new analysis/synthesis system of musical signals using prony's method-application to heavily damped percussive sounds," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 2053–2056.

[14] Jean Laroche and J-L Meillier, "Multichannel excitation/filter modeling of percussive sounds with application to the piano," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 329–344, 1994.

[15] Tuomas Paatero and Matti Karjalainen, "Kautz filters and generalized frequency resolution: Theory and audio applications," *Journal of the Audio Engineering Society*, vol. 51, no. 1/2, pp. 27–44, 2003.

[16] Adrien Sirdey, Olivier Derrien, Richard Kronland-Martinet, and Mitsuko Aramaki, "Modal analysis of impact sounds with esprit in gabor transforms," in *14th International Conference on Digital Audio Effects (DAFx-11)*, 2011, pp. 1–6.

[17] Tuomas Paatero and Matti Karjalainen, "New digital filter techniques for room response modeling," in *Audio Engineering Society Conference: 21st International Conference: Architectural Acoustics and Sound Reinforcement*. Audio Engineering Society, 2002.

[18] Esteban Maestre, Jonathan S Abel, Julius O Smith, and Gary P Scavone, "Constrained pole optimization for modal reverberation," in *Proc. of the 5th International Conference on Digital Audio Effects (DAFx)*, 2017.

[19] Adrien Sirdey, Olivier Derrien, Richard Kronland-Martinet, and Mitsuko Aramaki, "Esprit in gabor frames," in *Audio Engineering Society Conference: 45th International Conference: Applications of Time-Frequency Processing in Audio*. Audio Engineering Society, 2012.

[20] M Schoenle, N Fliege, and U Zölzer, "Parametric approximation of room impulse responses by multirate systems," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*. IEEE, 1993, vol. 1, pp. 153–156.

[21] Sahar Hashemgeloogerdi and Mark F Bocko, "Precise modeling of reverberant room responses using wavelet decomposition and orthonormal basis functions," *Journal of the Audio Engineering Society*, vol. 66, no. 1/2, pp. 21–33, 2018.

[22] Jean-Marc Jot, Laurent Cerveau, and Olivier Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," in *Audio Engineering Society Convention 103*. Audio Engineering Society, 1997.

[23] Richard Roy and Thomas Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 37, no. 7, pp. 984–995, 1989.

[24] Roland Badeau, Rémy Boyer, and Bertrand David, "Eds parametric modeling and tracking of audio signals," in *Proc. of the 5th International Conference on Digital Audio Effects (DAFx)*, 2002, pp. 139–144.

[25] Jean Laroche, "The use of the matrix pencil method for the spectrum analysis of musical signals," *The Journal of the Acoustical Society of America*, vol. 94, no. 4, pp. 1958–1965, 1993.

[26] Mathieu Lagrange and Bertrand Scherrer, "Two-step modal identification for increased resolution analysis of percussive sounds," in *Proc. Int. Conf. Digital Audio Effects (DAFx)*, 2008.

[27] Alan V Oppenheim, *Discrete-time signal processing*, Pearson Education India, 1999.

[28] NJ Fliege and U Zolzer, "Multi-complementary filter bank," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*. IEEE, 1993, vol. 3, pp. 193–196.

[29] Julius O Smith, "Audio fft filter banks," in *Proceedings of 12th International Conference on Digital Audio Effects (DAFx-09), Como*, 2009.

[30] Michael Goodwin, "Nonuniform filterbank design for audio signal modeling," in *Signals, Systems and Computers, 1996. Conference Record of the Thirtieth Asilomar Conference on*. IEEE, 1996, pp. 1229–1233.

[31] Mati Wax and Thomas Kailath, "Detection of signals by information theoretic criteria," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 387–392, 1985.

[32] Roland Badeau, Bertrand David, and Gaël Richard, "Selecting the modeling order for the esprit high resolution method: an alternative approach," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 2.

[33] Balázs Bank and Julius O Smith III, "A delayed parallel filter structure with an fir part having improved numerical properties," in *Audio Engineering Society Convention 136*. Audio Engineering Society, 2014.

[34] Jonathan S Abel and Patty Huang, "A simple, robust measure of reverberation echo density," in *Audio Engineering Society Convention 121*. Audio Engineering Society, 2006.

[35] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre, "webmushra: A comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.