
Article:
Wells, Jez (2021) Modal Decompositions of Impulse Responses for Parametric Interaction. Journal of the Audio Engineering Society. pp. 530-541. ISSN: 0004-7554

https://doi.org/10.17743/jaes.2021.0027

Modal Decompositions of Impulse Responses for Parametric Interaction

Jeremy Wells (Department of Music, University of York, UK.)

Abstract

A modelling system for the impulse responses (IRs) of reverberators is presented. The overarching purpose of this system is to offer similar levels of control over captured IRs to that of algorithmic reverberators whilst retaining their acoustic plausibility and, where desired, realism. Specifically, an approach to estimating the parameters of the model is presented which offers a significant reduction in the computational requirements of the matrix decomposition method ESPRIT, whilst offering vastly improved quality than is possible by using a single Fourier analysis. These methods are compared, first on large sets of short-duration synthetic signals, and then on a wide range of typical IRs, some many seconds in duration. Finally, systems that employ the model described and the analysis method it uses, are discussed.

1 Introduction

Reverberation, whether acoustically or synthetically generated, is a component of many listening experiences. Where those listening experiences are mediated by an audio production process, the reverberation(s) they contain are either directly captured, created by convolution of components with directly captured reverberation (usually as an impulse response) or created directly by passing components through networks of delay (which are connected in one or, more recently, two or more dimensions) and attenuation components [1, 2].

Control over reverberation is exercised by recording engineers and audio producers through selection and configuration of recording venue, selection and configuration of microphones relative to performers and the environment, and via synthetic means. The latter include mechano-electroacoustic (excitation of plates and springs etc.) and digital (via algorithmic and sampled artificial reverberators) convolution. Convolution of audio signals representing excitation sources with audio signals representing the reverberant response of the environment (sampled reverberation) offers a high level of realism, since the process is so closely analogous to those excitation signals occurring within that environment.

This convolution can be efficiently achieved with very low latency [3]. However, such samples of reverberation do not provide a model of the underlying processes that create them and, therefore, do not readily offer parameters by which they may be adjusted *post-hoc*. Algorithmic reverberators, on the other hand, do offer an explicit model of reverberation and, therefore, access to the parameters of that model for adjustment. However, such reverbs may not be able to closely mimic the sound of specific, or even generalised, real spaces due to limitations in the algorithms, which are often constrained in their complexity by design processes and the available computing resources available for them to run on.

Design methodologies have changed and computing resources have expanded exponentially, due to Moore's and other related laws, and the history of artificial reverberation is characterised primarily by these developments. However, both resource and design limitations remain issues currently for two- and three-dimensional networks of delays: at bandwidths that match or exceed that of human hearing the spatial and temporal sampling demands are still generally prohibitive for anything approaching real-time performance [4]. Additionally for this approach, there is the problem of creating a sufficiently detailed audit of a space's dimensions and materials (including their sound interacting properties). Where memory requirements can be accommodated, but execution times are too long for real-time, then the output of such a system can be rendered offline to an impulse response for subsequent use by a convolution engine.

With these issues in mind, a dichotomy can be observed between sampled reverberation applied via convolution and algorithmic reverb which is applied directly via the components it consists of. Provided the 'brute force' required for convolution can be provided then reverberation with features of any complexity can be applied, provided that a recording of it exists. However, adjustment of those features is not possible, since access to the original space or model is not available. Envelope modification to adjust reverberation time is possible [5] and at least one commercially available convolution processor offers a 'stretch' option, although this appears to be simple re-sampling of the impulse response [6]. Algorithmic reverberators on the other hand offer direct access to the parameters of the model, although that model may not offer sufficient or appropriate detail to match the level of realism, or other qualities, of sampled reverberation.

The work described here attempts to bridge the divide between algorithmic and sampled reverberation via a tool for the analysis, modification and resynthesis of reverberation signals. This is achieved via an audio model that is sufficiently specific to enable intuitive, plausible and meaningful modifications but general enough to accommodate the broad class of signals that can be reasonably described as reverberations. A spectral model is used which describes reverberations as the sum of sinusoids with exponentially changing amplitude which all have the same start point, but individual frequency, starting phase, starting amplitude and amplitude change rate. Modelling a signal as a set of exponentially decaying sinusoids is established and ongoing [7, 8]. The approach taken here is to use a very high number of components to the model the signal, offering very high objective and subjective fidelity to the original and allowing plausible and creatively useful variation of its parameters.

Modelling high-quality audio signals, possibly of many seconds in duration and with multiple channels, with an efficiency that is computationally tractable and an effectiveness that makes the models useful remains a significant challenge. Even though such an approach is inherently a model of the late reverberation, the large number components included also offers very good reconstruction of the temporal features of the early reflections in many impulse responses. This approach offers the convenience of being able to select an existing IR recording and then adjust it to taste or application.

Deriving a sufficiently detailed and flexible model and resynthesizing from modified versions of it is a computationally expensive task which, in many cases, is not tractable in real-time. This paper describes how the model parameters can be most effectively determined, and provides an example of how the model can be deployed in a studio-based reverb processing system, in order to achieve a

processing tool that can offer algorithmic-like interaction with sampled reverberation. A new modelling approach is presented that is a combination of spectral modelling techniques and relaxation. The paper is organised as follows: Section 2 provides a complete specification of the sound model, Section 3 compares three different methods (each with varying computational demands and qualities of outcome), Section 4 briefly describes a system that implements the model in a user-adjustable convolution reverb system and Section 5 summaries the key outcomes and achievements of this work.

2 Reverberation model

The system approximates a signal $s(t)$ which may be, but is not limited to, an impulse response as:

$$\tilde{s}(t) = \sum_{n=0}^{N-1} e^{a_n + j\varphi_n - t(\alpha_n + j2\pi f_n)} \quad (1)$$

where N is the total number of components in the model, a_n is the starting amplitude of the nth component in Nepers, φ_n is the starting phase, α_n is the exponential amplitude change rate and f_n is the frequency of the component. Since this is a digital system t represents a time index, where consecutive indices are spaced by a time period of $1/Fs$, where Fs is the sample rate of the system. This is referred to in the literature as the exponentially damped sinusoidal (EDS) model [9]. Typically, this model is deployed in frame-based analysis, where the signal is divided into sections within which its parameters are considered to be stationary. For audio signals that result from damped vibrational systems a model which incorporates exponential damping often allows a longer frame within which parameter stationarity can be assumed, particularly where frame boundaries are aligned with excitation points (i.e. note onsets).

An assumption in the application of the EDS model described here is that the four parameters do not change for the duration of the reverberation, which is deemed a reasonable choice since reverberation is considered the response of a time invariant damping system to an excitation input. In other words, the reverberation is analysed using a single frame. An implication of this, given the choice of parameters, is that whilst sinusoidal amplitude can change during the reverberation, its frequency does not. This greatly simplifies both analysis and resynthesis but retains relevance to most linear time-invariant reverberation (except that strongly characterised by swept components). A more sophisticated resynthesis system can accommodate modulations of these parameters, as discussed in the following sub-section, although this is at greater computational cost.

2.1 Model interactions

Once the analysis and modelling stage (described in detail in the next section) is complete, there is a set of four N-dimensional vectors $\mathbf{a}$, $\boldsymbol{\varphi}$, $\boldsymbol{\alpha}$ and $\mathbf{f}$ which represent N partials of the reverberation (alternatively, nodes of the vibrating structure that created it). Resynthesis can be achieved by a system that implements equation (1) above, or a Fourier-domain equivalent. The system described here uses a bank of oscillators, specifically the second-order digital waveguide resonator [10]. The output $y_n(t)$ of the nth resonator is:

$$y_n(t) = \begin{cases} 0, t < 0 \\ e^{a_n - \alpha_n t} x_n(t), t = 0,1.....T-1 \end{cases}$$

where

$$x_n(t) = \begin{cases} \cos(\varphi_n), t = 0 \\ \cos\left(\varphi_n + \frac{2\pi f_n}{F_s}\right), t = 1 \\ C x_n(t-1) - x_n(t-2), t = 2,3....T-1 \end{cases} \quad \text{and} \quad (2)$$

$$C = 2\cos\left(\frac{2\pi f_n}{F_s}\right)$$

Although we do not consider chirps within the scope of this paper, equation (2) can be adapted to incorporate linearly changing frequency with an additional recursion relation, based on the second order phase difference between samples, to update *C* at each iteration.

From this model a number of interactions with typical 'algorithmic' parameters are possible (many of those listed in [11], for example). In general: the reverberation time can be modified (in a frequency-dependent way, if required) by adjustment of the members of **α**, the room size and modal density can be modified by redistribution of the members of **f**, the spectral envelope can be altered via **a**. Although this a late reverberation model, the inclusion of start phase (within **φ**) and the number of model components, *N* (which can be as high as *T*/4, as discussed later) retains the temporal structure of the early reflections of impulse responses.

2.2 Accounting for air absorption

As sound waves propagate through air in typical atmospheric conditions there is some attenuation due to viscous losses. The amount of energy loss increases with frequency. In many acoustic scenarios, particularly for shorter distances (such as those travelled between sound source and listener - the direct sound – and early reflections) this effect is negligible. However for large, highly reverberant spaces it has a significant effect on the quality of sound. This 'darkening' which happens over the duration of a long reverberation is an important feature of the responses of such spaces. This attenuation (presented here as 'per sample' rather than 'per metre') is described by the Bass formula [12]:

$$\alpha_{\text{air},n} = \frac{c}{F_s} f_n^2 \left(1.84 \times 10^{-11} \left(\frac{Y}{Y_0}\right)^{\frac{1}{2}} \right) + \left(\frac{Y}{Y_0}\right)^{\frac{-5}{2}} \left(\left(\frac{0.01278 e^{\frac{-2239.1}{T}}}{f_{r,O} + \left(\frac{f_n^2}{f_{r,O}}\right)} \right) + \left(\frac{0.1068 e^{\frac{-3352.0}{T}}}{f_{r,N} + \left(\frac{f_n^2}{f_{r,N}}\right)} \right) \right) \quad (3)$$

where *c* is the speed of sound in air, *Y* is the temperature in degrees Kelvin and Y_0 is '293.15 K … the reference temperature' [12], and f_{oe} and $f_{r,\mathrm{N}}$ are the scaled relaxation frequencies for oxygen and nitrogen respectively and are given by equations (1) and (4) to (6) in [12].

When applying large time scaling factors to reverberation sampled from an acoustic space, the quality of the scaled version is significantly dissimilar to that which would be expected of a more reverberant version of that space: there is an unnatural brightness that persists throughout the duration of the reverberation. This is because the additional air absorption implied by the longer reverberation time is not being considered. This can be ameliorated by making the following modification to the time scaling of the amplitude change:

$$\tilde{\alpha}_n = \alpha_{\text{air},n} + \gamma\alpha_{\text{surfaces},n} \text{ where}$$
$$\alpha_{\text{surfaces},n} = \alpha_n - \alpha_{\text{air},n} \quad (4)$$

and $\alpha_{\text{air},n}$ is the attenuation per sample given by equation (4), α_n is the estimate of the total attenuation per sample, which is assumed to be the sum of attenuation due to absorption in air and at interactions with surfaces within the space.

3 Estimation of model parameters

The success of a model-based sound modification system is dependent upon the suitability of the model to the sounds being modified, and the suitability of the model's parameters to the intended modifications and the accuracy of the estimation of those parameters. In this section three alternative approaches to estimating the parameters are described and compared, including a novel method, referred to as 'modelled pursuits'. Each approach has different computational demands in terms of required memory and number of calculations. The relative performance of the three approaches on a variety of impulse responses is presented at the end of this section. Each estimate is derived from the time-domain impulse response *s*(*t*).

3.1 Estimation of signal parameters via rotational invariance techniques (ESPRIT)

The ESPRIT method was first described for direction of arrival (DOA) estimation of narrowband signals arriving at an array of antennas [13] in the presence of uncorrelated (white) noise. A constraint on the array is that it contains pairs of elements with identical properties, with all pairs separated by a fixed displacement. Between the elements of each pair the parameters (DOA in this case) are assumed to be constant for each signal source, with only a phase delay between the two elements in each pair due to the displacement. This approach has been adapted to estimating the parameters of signals comprising exponentially decaying sinusoids [14]. The method is also described in detail in [15] and [16] and the reader is directed there for a detailed treatment, but a summary is given here.

In this approach a signal subspace is estimated which is a matrix of complex poles representing sinusoids with exponential amplitude change. This subspace is estimated by the singular value decomposition of the Hankel matrix **S** of the time-domain samples of the signal:

$$\mathbf{S} = \begin{bmatrix} s[0] & s[1] & \dots & s\left[\left\lfloor \frac{T}{2} \right\rfloor - 1\right] \\ s[1] & s[2] & \dots & s\left[\left\lfloor \frac{T}{2} \right\rfloor\right] \\ \vdots & \vdots & & \vdots \\ s\left[\left\lfloor \frac{T}{2} \right\rfloor - 1\right] & s\left[\left\lfloor \frac{T}{2} \right\rfloor\right] & \dots & s[T] \end{bmatrix} \quad (5)$$

The SVD of this matrix yields a matrix **U** containing basis vectors that span the signal subspace. If these vectors are assumed to be sinusoids with exponentially changing amplitude, then a single sample shift for each vector can be achieved by a phase shift, that represents the time displacement and a multiplier that represents the amplitude change that occurs over that time. Both the phase shift and multiplier can be combined into a complex exponential, which is the pole that describes the frequency and decay rate of the sinusoid. If the sinusoids are each considered to be decaying at exponential rates and the sampling rate is constant then the pole describing each sinusoid does not vary from sample to sample, which is the invariance property that this estimation method exploits. Therefore the poles are estimated as the diagonal of:

$$\Phi = \left(U^{\downarrow}\right)^{+} U^{\uparrow} \quad (6)$$

where + indicates the pseudo-inverse operator and the upward arrow indicates a version of the matrix with the first row removed and the downward arrow indicates a version with the last row removed. Essentially, (6) determines the phase shift and amplitude scaling of each pole by division of one matrix by the same matrix shifted forwards by one sample.

Once the poles have been estimated they are used to form a Vandermonde matrix. Multiplication of the pseudo-inverse of this matrix with the original signal vector yields estimates of the starting amplitude and phase of each sinusoid which minimise the mean squared error between the original signal and that synthesized from the model. Equation (6) assumes that the matrix is square. Although this is not a necessity for ESPRIT or SVD more generally, in this specific application the matrix is square and its width and height are equal to half the length of the signal being analysed. A maximum of $T/4$ poles (and therefore $T/4$ overall model components) can be recovered.

This algorithm has complexity $O(T^3)$ and for long impulse responses (e.g. lasting multiple seconds) can make vast demands on memory (since a square matrix must be formed with width and height both equal to half the signal length in samples) and processing. For this reason, it is often used on critically sampled sub-band representations (e.g. [11]).

3.2 Estimation from single discrete-time Fourier transform

The discrete-time Fourier transform (DFT) is a ubiquitous spectral analysis tool in many areas of discrete time series analysis. Many efficient (often referred to as ‘fast’ Fourier transforms) implementations exist which are relatively computationally un-demanding and can now easily be calculated in real-time on even very modest computing systems. The DFT is an orthogonal transform which yields the parameters (stationary amplitude and starting phase) of sinusoidal atoms at integer

multiples of a fundamental frequency given by F_s/K where K is the length of the DFT output. In the uses of the DFT described in this work, zero-padding is implemented to improve the selection of peaks and the amount of spectral detail around those peaks. Here 8x padding is used (where T is rounded down to the nearest power-of-2), formally expressed as:

$$K = 2^{\lfloor log_2 8T \rfloor} \tag{7}$$

The DFT is an $O(T\log(T))$ algorithm.

3.2.1 Frequency estimation

Although the DFT does not directly yield fine frequency estimates of underlying components, a number of methods exist to derive these. Many of these require additional DFTs (e.g. for frequency reassignment, the derivative method or the phase vocoder) but a similar level of accuracy is possible via parabolic interpolation of the log magnitude spectrum of a single DFT of the signal [17]. The frequency, f, of a sinusoidal component is estimated from:

$$\begin{aligned} f_n &= \frac{F_s}{K}\left(k_n + \kappa_n\right), \text{ where} \\ \kappa_n &= \frac{1}{2}\left(\frac{\ln\left|S_{k_n-1}\right| - \ln\left|S_{k_n+1}\right|}{\ln\left|S_{k_n-1}\right| - 2\ln\left|S_{k_n+1}\right| + \ln\left|S_{k_n+1}\right|}\right) \end{aligned} \tag{8}$$

$\boldsymbol{S}$ is the DFT of the time-domain signal $\boldsymbol{s}$, k_n is the index of nth peak in $\boldsymbol{S}$ with k=0,1,2,..., $\lfloor K/2 \rfloor + 1$ due to the complex conjugate symmetry of the DFT of real signals.

3.2.2 Phase estimation

Phase can be interpolated directly from the DFT:

$$\varphi_n \cong \arg\left(S_{k_n}\right) + \frac{\kappa_n\left(\arg\left(S_{k_n+1}\right) - \arg\left(S_{k_n-1}\right)\right)}{2} \tag{9}$$

3.2.3 Amplitude decay rate estimation

Although the DFT decomposes a signal into sinusoids that are stationary in terms of amplitude, rates of intra-frame amplitude change can be estimated from the derivative of the phase with respect to frequency at magnitude peaks in the DFT spectrum. Where no tapering is applied to the signal (i.e. for a rectangular window) the relationship between the phase derivative and the amplitude change rate is [18]:

$$\frac{d\left(arg\left(S_{k_n}\right)\right)}{df} = 2\pi\left(\frac{1}{\xi_n} + \frac{1}{(1 - e^{\xi_n})} - 1\right) \tag{10}$$

where $\xi_n = -\alpha_n / T$. This function is shown in Figure 3 for values of ξ between 0 and +/-15, which correspond to between 0dB and 130 dB of total amplitude reduction. Beyond this range of values, the modelled components become increasingly impulsive (impulse at the very start of the signal for large negative ξ, impulse at the very end of the signal for large positive ξ). For very large negative values of ξ it tends to 0, for large positive values it tends to -2π.

It is not possible to rearrange (10) so that ξ can be directly determined so a look-up table (LUT) is used. At initialisation, a 100 000 point table with equidistantly spaced values of ξ between zero and 500 is calculated. From this initial table a much smaller table (100 points) is derived using piece-wise

cubic spline interpolation is used, with equidistantly spaced values of phase derivative in the range 0 to 2π. This smaller look-up table is then used for estimation of the phase derivative from the measured ξ, and the larger table is discarded. Because the phase derivative values are linearly spaced this table look-up can be performed by direct addressing which, combined with the small table size, greatly speeds up the estimation of ξ. The phase derivative is estimated from the phase difference across a peak as

$$\frac{d\left(\arg\left(S_{k_n}\right)\right)}{df} \cong \frac{K\left(\arg\left(S_{k_n+1}\right)-\arg\left(S_{k_n-1}\right)\right)}{2T} \tag{11}$$

With a zero-padding factor of 8, equation (10) gives a good approximation of the phase derivative (as was shown to be the case for the Hann window in [19]). This measured value is then used with the LUT to find, via piece-wise cubic spline interpolation, the corresponding value of ξ_n. Prior to the calculation of equation (11) the phases, arg(**S**) are unwrapped such that they are monotonically decreasing with increasing index.

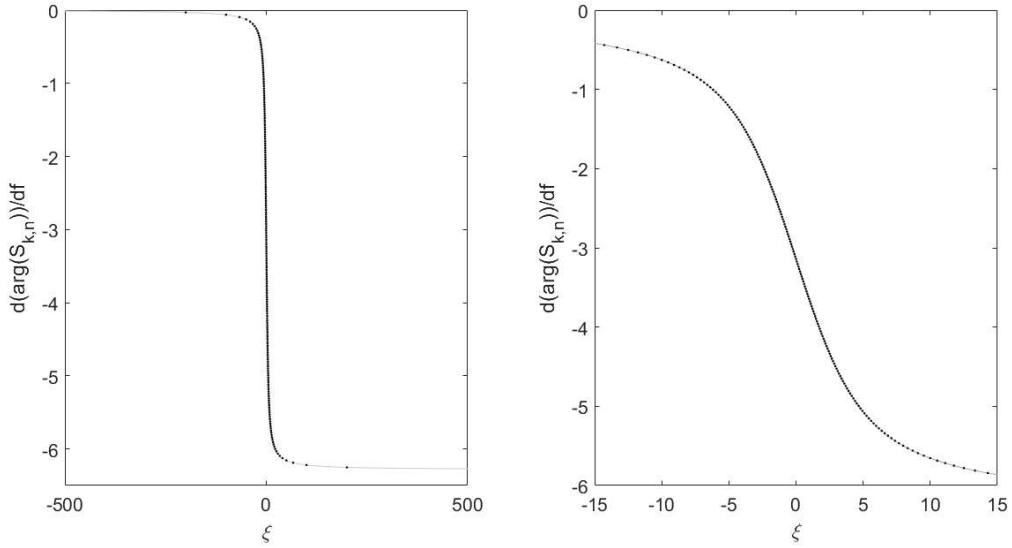

Figure 1: Effect of amplitude change amount on DFT peak phase. The right panel is a zoomed version of the left. The grey line is the underlying continuous function, the black dots indicate sample points that make up the linear (directly indexed) LUT.

3.2.4 Amplitude estimation

Since the rate of amplitude decay determines the spread of energy across bins in the DFT the starting amplitude of the component can only be estimated once α_n has been found. Firstly, the magnitude of the DFT at the peak is estimated via parabolic interpolation:

$$\left|S_{k_n}\right| = e^{\left(\ln\left|S_{k,n}\right| - \frac{\kappa_n}{4}\left(\ln\left|S_{k,n-1}\right| - \ln\left|S_{k,n+1}\right|\right)\right)} \tag{12}$$

Then, from this, the true amplitude is estimated, using knowledge of the Fourier transform of the window function. For a rectangular window (i.e. when no tapering is applied) this is given by [18]

$$a_n = \begin{matrix} \ln\left(\frac{|S_{k_n}|\xi_n}{e^{\alpha_n}-1}\right), \xi_n \neq 0 \\ |S_{k_n}|, \xi_n = 0 \end{matrix} \quad (13)$$

3.3 Iterative estimation from repeated DFTs ('modelled pursuits')

The presence of multiple components within a Fourier spectrum can lead to biased parameter estimates. Where components are well localised and isolated in frequency the effect of the bias will be small. However, fast-decaying elements will be widely spread in the Fourier domain and typical impulse responses (e.g. those captured in rooms) contain a large number of room modes whose density increases with frequency. In such cases the effect of bias on parameter estimates and, therefore, on the model as a whole are likely to be significant. In particular the presence of particularly energetic modes will hamper accurate estimation of weaker components. Whilst windowing of signals will reduce biasing due to spectral leakage, it comes at the cost of blurring of components, which hampers the identification of separate model components.

This bias can be reduced by an iterative process which estimates the parameters of the highest magnitude component in the DFT, synthesizes the component in the time domain from those estimates and subtracts it from the signal, leaving a residual. This residual is then analysed in the same way (requiring a new DFT to be calculated): the component with highest magnitude in the residual spectrum is estimated, synthesized in the time domain and subtracted to yield a new residual. This process is continued until a stopping criterion is reached. This might be a threshold value for the mean squared error (MSE) between the input and modelled signal, a point at which the MSE ceases to monotonically decrease, or a limit to the number of components allowed within the model output. Provided their parameter estimates are accurate, components removed at each stage will not bias subsequent analysis stages.

This approach bears some similarity to matching pursuit (MP) algorithms and Spectral Modelling Synthesis (SMS) [20]. In the original MP algorithm described by Mallat and Zhang [21] a signal is decomposed into the linear sum of a set of vectors obtained from a redundant dictionary. This decomposition is achieved by an iterative approach. Firstly, the inner product of the signal with all dictionary elements is calculated and the element with the highest inner product is chosen as the initial atom in the decomposition.

$$\mathbf{s} = \langle \mathbf{s}, \mathbf{g}_{\boldsymbol{\gamma_0}} \rangle \mathbf{g}_{\boldsymbol{\gamma_0}} + \mathbf{R^0 s} \quad (14)$$

where **s** is the signal at all time instants, $\mathbf{g}_{\boldsymbol{\gamma}}$ is a dictionary element (of index $\boldsymbol{\gamma}$, with $\boldsymbol{\gamma_0}$ the index of the zeroth chosen atom) and $\mathbf{R^0 s}$ is the residual after approximating **s** with the zeroth index atom. The process is repeated on the residual to yield a two-element decomposition:

$$\mathbf{s} = \langle \mathbf{s}, \mathbf{g}_{\boldsymbol{\gamma_0}} \rangle \mathbf{g}_{\boldsymbol{\gamma_0}} + \langle \mathbf{s}, \mathbf{g}_{\boldsymbol{\gamma_1}} \rangle \mathbf{g}_{\boldsymbol{\gamma_1}} + \mathbf{R^1 s} \quad (15)$$

This process continues until a stopping point is reached (such as the energy of $\mathbf{R^n s}$ reducing to a specified threshold, or after a certain number of iterations). $\mathbf{R^n s}$ approaches 0 with increasing *n*.

In the approach introduced in this section the same iterative procedure is followed, but the dictionary search (finding the inner product of the signal residual with each of a set of fixed atoms) is replaced by the modelling of an atom using parameter estimates derived (as described in section 3.3) from the DFT. For this reason, the approach described here is referred to as 'modelled pursuits' (MoP). It bears some similarity to the relaxation-based estimation of Liu et al. [22], however MoP directly estimates parameters. It does not, for example, undertake one-dimensional error minimisation to find the amplitude decay rate.

There are two possible approaches to estimating the amplitude of the atoms. The first, adopting the method of MP, finds the inner product of the residual and the energy-normalised atom. The second, estimates the amplitude directly from the Fourier spectrum (i.e. using eqns. (14) and (15)). The relative effectiveness of both methods is compared in the experiments described later in this section.

The iteration of this algorithm is stopped when any of the following criteria are met:

1. The energy in $\mathbf{R}^n\mathbf{s}$ becomes greater than $\mathbf{s}$. This criterion should not be met when the inner product method is used to estimate component amplitude, since residual energy decreases monotonically with MP. However, this situation can occur with direct estimation of amplitude particularly if a side lobe in the Fourier domain is mistaken for a main lobe, since this can cause extreme values of amplitude change rate to be estimated.

2. The energy in $\mathbf{R}^n\mathbf{s}$ reaches -96 dB relative to the energy in the original signal $\mathbf{s}$. In practice this is unlikely, but it represents the entire theoretical dynamic range of a 16 bit system.

3. The number of iterations n reaches $T/4$, where T is the total length of the input signal in samples. This ensures that the storage required for the model parameters (of which there are four per decaying sinusoid: amplitude, phase, frequency and decay rate) does not exceed that originally required by the time-domain samples of the impulse response being modelled. This also matches the maximum number of components that can be estimated using the ESPRIT method (summarised in Section 3.1).

Since it is often the third stopping criterion that is reached first, MoP can be considered an $O(T^2\log(T))$ algorithm, placing it between single-frame DFT-based estimation and ESPRIT in computational complexity.

3.4 Comparison of estimation methods

Having presented two new methods, along with an established approach for estimating the parameters of eqn. (1), their relative effectiveness is now compared. This is done in two ways. The first test is designed to assess the general ability of the techniques to successfully extract the component type (damped sinusoids) from a signal. This is intended to enable a straightforward comparison between these approaches in terms of how well they identify and parametrise signal components in differing levels of noise Specifically, the various methods described are tested with short (2000 sample) synthesized frames, comprising sinusoids with exponential amplitude change and white Gaussian noise. This enables their performance at specific signal-to-noise ratios (SNR) to be compared. The second test is intended to illustrate their relative performance on specific

examples of the intended signal type. For this they are tested on 20 impulse responses (18 are acoustic, two are synthetic) which represent a range of different reverberation types and qualities.

3.4.1 Short frames comparison

As previously noted, ESPRIT is an extremely demanding estimation process for long frame sizes. Therefore, to enable a sufficient number of trials to produce reliable results a signal length for this trial of *T* = 2000 samples is chosen. Each trial compares the signal-to-noise ratio (SNR) of a modelled signal (with estimates derived from either of the three methods) with the SNR of the original signal from which it is derived, which is set of sinusoids combined with white Gaussian noise (WGN) of a specified relative level. The number of sinusoidal components *N* in each test signal is randomly selected from a uniform probability distribution $N \sim U(1, \lfloor T/4 \rfloor)$. For each sinusoidal component in a signal the amplitude is selected using the uniform probability distribution U(0,1), the frequency from U(0,Fs/2), the amplitude change from U(-96,96) dB and the phase from U(-π, π). For each test signal the set of sinusoids for that signal are created (with parameters drawn from the probability distributions described above) and summed. WGN is then created, scaled and added to the sinusoids so that a 'sinusoids plus noise' signal of the desired SNR is created.

The trial compares input and output SNR at input SNRs of -40, -20, 0, 20, 40, 60, 80 and 100. At each of these input SNRs 10 000 individual signals are synthesized for comparison. They are each modelled via the following estimation methods:

1. ESPRIT
2. MoP (rectangular window, direct amplitude estimation)
3. MoP (rectangular window, inner product amplitude estimation)
4. Single-DFT (rectangular window, direct amplitude estimation)
5. Single-DFT (rectangular window, inner product amplitude estimation)

For each analysis no attempt is made to determine the model order and it is assumed to be $\lfloor T/4 \rfloor$ in each case. The sample rate is 44.1 kHz.

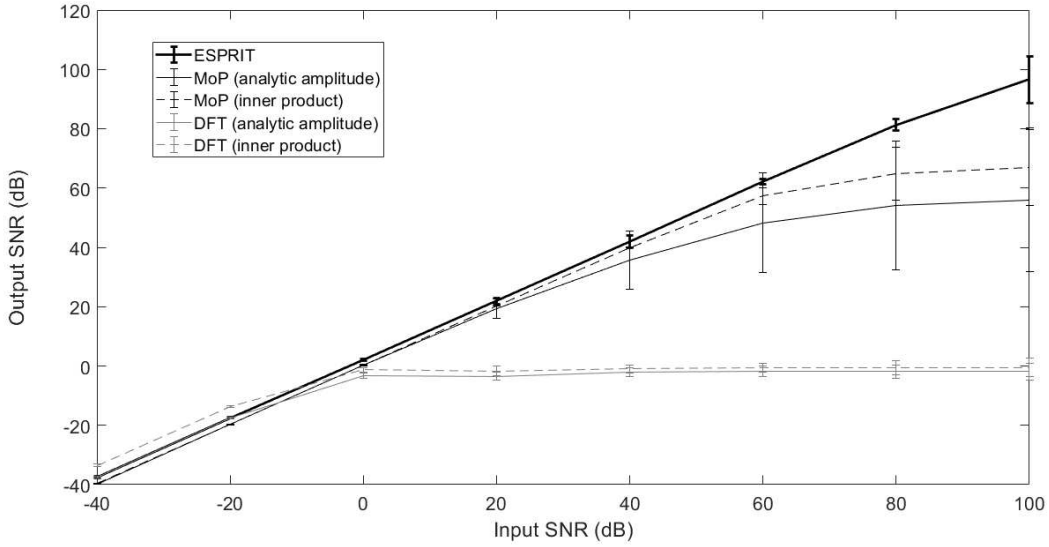

Figure 2: Mean output SNR versus input SNR (vertical bars indicate the standard deviation) for decomposition methods, T = 2000.

Fig. 2 shows the relationship between input and output SNR (the mean of 10 000 values at each data point). It can be seen that ESPRIT performs very well, with close to equivalence between input and output SNR and is the best estimator at medium input SNRs. Inner-product MoP offers performance close to ESPRIT except at high input SNRs (30 dB poorer output SNR at an input SNR of 100 dB), with analytic MoP about 10 dB worse than that at the highest input SNR. Both single-DFT approaches exhibit a ceiling of 0 dB output SNR although, perhaps surprisingly, the inner-product variant performs better than all methods for the lowest input SNRs.

Fig. 3 shows the execution time (the mean of 80 000 values, from all data points). Although execution time is dependent on implementation and specific hardware, these measures do give some indication of the scale of the different demands for each of estimation methods. The ESPRIT estimation uses the HR_Analysis function from the DESAM toolbox, modified to use a faster singular value decomposition function [16, 23]. Other estimation methods use a Matlab executable (MEX) implementation of the waveguide oscillator of eqn. (2) in order to minimize the time taken for the synthesis of sinusoids. For the application described in Section 4 of this paper, the modelling is performed offline prior to deployment in a convolution reverb.

If the five methods are ranked according to output SNR and execution time, there is an exact correspondence between those two rankings (e.g. ESPRIT is the best performing but is also the most computationally demanding).

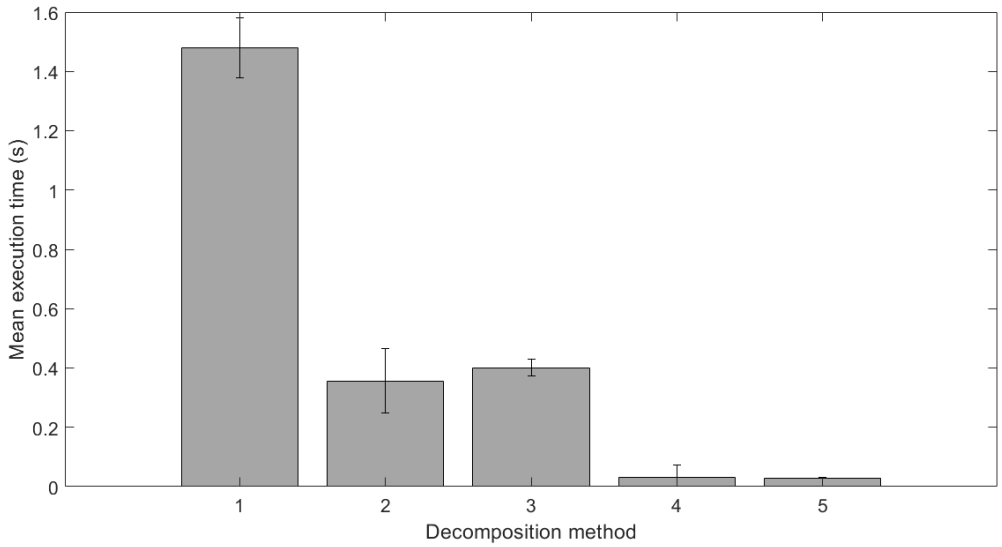

Figure 3: Relative execution times for decomposition methods. These are (1) ESPRIT, (2) MoP (direct amplitude estimation), (3) MoP (inner product amplitude estimation), (4) single-DFT (direct amplitude estimation), and (5) single-DFT (inner product amplitude estimation).

3.4.2 Impulse responses comparison

The algorithms are tested on eleven varying IRs of which all but three are freely available from the *Open Acoustic Impulse Response Library* [24]. The majority of these are two-channel, giving a total of 20 IR signals in total. They are described in Table 1. They are chosen to represent an accessible collection of IRs that vary in terms of origin, quality (many are captured at high SNRs by using the swept-sine method, but there is also a 'balloon pop' capture) and duration. For each IR the same five decomposition methods are tested and compared.

The IR recordings are used 'as is', although they are converted to a sample rate of 44.1 kHz where necessary and they are automatically edited to ensure that there is no silence (or near-silence) at the start of the files before the IRs begin. This automatic truncation of the file starts is achieved via a straightforward process: the reverberation time (RT_{60}) is estimated and from this a single exponentially decaying envelope of that duration is synthesized and convolved with the original audio. The square of the first-order difference of this signal is then found and peak-normalised. The start of the IR is taken as the sample at which this signal exceeds a threshold (here set to -90 dB_{FS}).The purpose of the convolution step is to accentuate the onset of the IR: the point at which the similarity between the synthesized envelope and the IR is greatest should also be the point at which their front edges coincide, emphasising the onset.

A difference between the testing in this and the previous section is that the ESPRIT method is used on sub-banded versions of the IRs. An eight band decomposition with perfect reconstruction is used, except in one case of a very long impulse response where a 16 band decomposition is used to make the computation tractable within 32 GB of RAM. Again, the implementation of ESPRIT and the sub-band decomposition in the DESAM toolbox [16] is used in these comparisons

The results for each of these estimation methods are shown in Table 2. This shows the following for each method: the residual to signal ratio (RSR) ς in dB to one decimal place, the number of terms (components) in the derived model (equal to *T*/4 for ESPRIT, up to *T*/4 for the other methods) and the objective grade descriptor (OGD), rounded to the nearest integer. The RSR is given by:

$$\varsigma = 10\, log_{10} \frac{\sum_t (s(t) - \tilde{s}(t))^2}{\sum_t s(t)^2}, t \in \mathbb{Z}^{\geq} \tag{16}$$

The OGD is a computed measure of perceptual quality described in ITU-R Recommendation BS.1387, *Method for Objective Measurements of Perceived Audio Quality* and implemented in a Matlab toolbox [26]. The descriptors for the grades are: imperceptible difference between the original and modelled signal (0), a perceptible but not annoying difference (-1), a slightly annoying difference (-2), an annoying difference (-3) and a very annoying difference (-4).

For MoP it can be seen in some cases the number of terms does not reach *T*/4. This is where an atom has been selected that leads to an increase in the energy remaining in the residual, which is one of the stopping criteria. This only occurs for the direct amplitude estimation method, as expected.

IR name	Description	Provenance	Channels
NCEM	Blumlein pair capture derived from a first-order ambisonic capture from the National Centre for Early Music (NCEM), York.	Open AIR [24] (receiver position 2)	2
480L Hall	Captured from Lexicon 480L reverberator , 'small hall' algorithm.	Hopkins IR library [25] ('480L Halls 05 Small Hall')	2
EMT Plate	Captured from EMT 'bright' plate.	Hopkins ('EMT 03 Bright Plate 03')	2
Kiln	Captured in Errol Brickworks kiln.	Open AIR	2
Koli in Summer	Captured in Koli national park, Finland.	Open AIR	2
Koli in Winter	Captured in Koli national park, Finland.	Open AIR	2
Lounge 1	Domestic living room.	Open AIR	1
Perth Hall	Captured on the balcony of Perth City Hall, UK.	Open AIR	2
York Minster	Captured in the nave of York Minster, UK.	Open AIR	2
Lounge 2	Captured in the author's living room.	Recording by author	1
Cresswell	Captured in Cresswell Crags, Derbyshire, UK.	Open AIR	2

Table 1: Impulse responses used in analysis

IR name	ESPRIT RSR (dB)	terms	ODG	MoP (direct) RSR	terms	ODG	MoP (inner product) RSR	terms	ODG	single DFT (direct) RSR	terms	ODG	single DFT (product) RSR	terms	ODG
NCEM L	-50.1	24121	0	-49.7	24121	0	-48.8	24121	0	0.9	5074	-4	0.5	5074	-4
NCEM R	-50.3	24121	0	-50.0	24121	0	-49.2	24121	0	0.7	4488	-4	0.3	4488	-4
480L Hall L	-48.1	19280	0	-36.8	19280	0	-34.3	19280	0	1.4	5500	-3	0.8	5500	-3
480L Hall R	-47.5	19280	0	-36.3	19280	-1	-33.8	19280	-1	2.5	5545	-3	1.0	5545	-3
EMT Plate L	-49.8	53487	0	-63.8	53487	0	-63.3	53487	0	0.8	12772	-4	0.4	12772	-4
EMT Plate R	-49.2	53487	0	-61.5	53487	0	-61.0	53487	0	0.7	12504	-4	0.4	12504	-4
Kiln L	-49.0	5442	-1	-32.3	2175	-1	-47.1	5442	0	1.6	1682	-2	0.8	1682	-2
Kiln R	-49.4	5442	-1	-28.8	1609	-1	-47.8	5442	0	4.6	1620	-3	2.5	1620	-3
Koli in Summer L	-42.3	8695	0	-53.0	8695	0	-51.9	8695	0	0.8	2146	-1	0.4	2146	-1
Koli in Summer R	-46.4	8695	0	-55.2	8695	0	-53.7	8695	0	0.7	1946	-1	0.4	1946	-1
Koli in Winter L	-43.0	7037	0	-51.4	7037	0	-50.3	7037	0	2.5	1990	-1	2.4	1990	-1
Koli in Winter R	-47.1	7037	0	-52.2	7037	0	-50.9	7037	0	1.5	1957	-1	1.5	1957	-1
Lounge 1	-25.8	9857	-2	0.0	0	N/A	-57.5	9857	-2	0.1	1663	-4	0.0	1663	-4
Perth Hall L	-48.1	55433	0	-55.4	55433	0	-54.8	55433	0	0.8	27786	-4	0.3	27786	-4
Perth Hall R	-48.0	55433	0	-55.6	55433	0	-55.1	55433	0	0.8	27840	-4	0.4	27840	-4
York Minster L	-49.2	109481	0	-54.2	109481	0	-53.5	109481	0	0.8	44770	-4	0.5	44770	-4
York Minster R	-47.9	109481	0	-53.8	109481	0	-53.1	109481	0	0.8	45155	-4	0.4	45155	-4
Lounge 2	-50.9	7724	0	-45.2	6203	0	-48.6	7724	0	2.0	1426	-1	1.1	1426	-1
Cresswell L	-51.9	20648	0	-58.5	20648	0	-57.3	20648	0	0.6	4567	-4	0.3	4567	-4
Cresswell R	-51.4	20648	0	-55.8	20648	0	-54.6	20648	0	0.7	5127	-4	0.3	5127	-4

Table 2: Residual-to-signal ratio (RSR), number of terms in model and objective difference grade (ODG) for the five estimation methods

There are three overall observations that are drawn from this set of data.

1. Single-DFT analysis does not perform well enough to produce plausible and robust models of reverberation for high-quality audio applications. According to the ODG measure there is always a perceptible difference, and this difference is often 'annoying; or 'very annoying'. Whereas ESPRIT and MoP offer IR models that in most cases, when resynthesized, are perceptually indistinguishable from, or very closely resemble, the original audio input, this is not the case with single-DFT analysis. Typically, the gross temporal structure is successfully captured but the IR is noisy and often excessively bright.

2. MoP has a performance that is on a par with sub-banded ESPRIT, although there is substantial variation across individual cases with ESPRIT sometimes substantially out-performing MoP, MoP sometimes substantially out-performing ESPRIT and the two being on a par in other cases. Because of the significant relative (and often prohibitive) expense of ESPRIT, even when sub-banded, MoP is a good choice for creating the model described in Section 2.

3. Analytic MoP performs better than the inner-product version except where the process terminates early due to the residual energy suddenly rising above that of the original input signal. This is surprising given the results in the previous sub-section. An explanation is that the analytic method derives its estimates using data at a spectral peak, whereas the inner-product method estimates amplitude by comparison of the atom with the global signal. It may be that, in the latter case, energy that is actually due to other atoms, or noise, is being erroneously included, making the process less effective at later iterations. That said, there are some catastrophic early terminations for the analytic method (e.g. 'kiln', and one instance, 'lounge 1', where it completely fails) and more consistent modelling quality is obtained with the inner-product method.

4 Interactive Reverberation Modelling System

Because of the desirable trade-off between modelling quality and computational demands, demonstrated in the previous section, MoP has been adopted for an interactive reverberation modelling developed by this author which is available as a software demonstrator [27]. Models are created from audio files containing IRs. If the parameters of the model are changed a new version of the IR is resynthesized in as close to real-time as possible and then used in a convolution reverberator. Strategies to ensure fast resynthesis are not a focus of this paper, but they can include sub-banding/decimation-in-time in conjunction with the waveguide oscillator described by eqn. (2).

The software demonstrator offers user control over the trade-off between resynthesis time and quality, and measures and displays the computation time at each IR resynthesis. This modelling system has also been incorporated into a commercially available digital reverberator for studio-based applications [28]. In these systems IR modelling takes place offline and real-time processing is achieved by convolution with audio synthesized from the model. Each time a reverberation parameter is adjusted the IR is resynthesized. Even for large IRs this resynthesis can be performed in faster-than-real-time on modest hardware if decimation and sub-banding are used in conjunction with the waveguide oscillator.

The demonstrator offers control over reverberation time and density, and room size. Once a model of the form of eqn. (1) has been generated, these parametric controls are straightforwardly

implemented. Reverberation time changes are achieved by modification of the decay rate parameter of each atom. This modification is frequency-dependent, in order to take account of air-absorption in accordance with eqn. (3). Modal density can be reduced by selectively removing atoms from the model. The possibility of increasing the modal density is offered by creating a set of 'shadow' atoms, which are replicas of each of the model components, with their frequencies scaled by $\sqrt{0.5}$ (this value is chosen as it is an irrational number which is the geometrical mean between two octaves and, as such, in order to reduce the coincidence in frequency of the original and new set of modes). As the modal density is increased above 100 % more of these shadow components are added into the model. Room size is varied by scaling atom frequencies according to

$$\widetilde{f_n} = f_n 2^{-r\left(\frac{F_s - 2f_n}{F_s}\right)} \tag{17}$$

where *r* is the base-2 logarithm of the room size multiplier. This gives minimal downward shift at high frequencies (none at Nyquist) and greatest shift at lowest frequencies (it is the reciprocal of the room size multiplier at 0 Hz).

Demonstrations of these modifications can be heard at a supporting website for this paper [28]. A video showing how the demonstrator can be used, including the effects of these types of parameter changes on the model (and the computation time for resynthesis of the model) is available [29]. Although these modifications are simply conceived they are effective in changing perception of density and room size when used in conjunction with a high-quality model derived using the methods described previously in the paper. Other modifications typically available in algorithmic reverberators are also straightforward to implement via this model.

5 Summary

This paper has presented an approach to high-quality flexible reverberation modelling, which aims to enable algorithmic control of existing impulse responses. In particular, it has described how parameters of the model can be directly estimated from single or iterative use of a zero-padded DFT, and how those modelling approaches compare with a high-quality, but computationally intensive, matrix decomposition method, ESPRIT. The iterative approach, MoP, has been shown to offer reasonable quality compared to ESPRIT for a considerable computational saving. In particular, MoP combines tractability with high quality when used on actual IR signals. Finally, an example of the deployment of MoP in an interactive reverberation modelling system, with simple but effective approaches to algorithmic interaction, has been given.

6 References

1. V. Välimäki, J.D. Parker, L. Savioja, J. Smith, and J. Abel, "Fifty Years of Artificial Reverberation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 2, pp. 1421-1448 (July 2012). https://doi.org/10.1109/TASL.2012.2189567

2. V. Välimäki, J.D. Parker, L. Savioja, J. Smith, and J. Abel, "More Than Fifty Years of Artificial Reverberation," presented at the *60th International Conference of the Audio Engineering Society* (2016 January), paper K-1.

3. W. G. Gardner, "Efficient Convolution Without Input-Output Delay". *J. Audio Eng. Soc*, vol. 43, no. 3, pp. 127-136 (1995 March).

4. F. Pind, C. Jeong, A. Engsig-Karup, J. Hesthaven, and J. Strømann-Andersen, “Time Domain Room Acoustic Simulations Using the Spectral Element Method,” *Journal of the Acoustical Society of America*, vol. 145, no. 6, pp. 3299-3310 (2019 June). https://doi.org/10.1121/10.0002448
5. J. Wells, M. Beeson and D. Murphy, “Temporal Matching of 2D and 3D Wave-Based Acoustic Modeling for Efficient and Realistic Simulation of Rooms,” presented at the *126th Convention of the Audio Engineering Society* (2009 May), paper 7697.
6. Christian Knufinke Software. “SIR version 2.24 Manual”. http://www.siraudiotools.com/downloads/SIR2_Manual.pdf (accessed May 9, 2021).
7. J. Abel, S. Coffin and K. Spratt, “A Modal Architecture for Artificial Reverberation with Application to Room Acoustics Modeling,” presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), paper 9208.
8. L. Birnie, T. Abhayapala, H. Chen, and P. Samarasinghe, “Sound Source Localization in a Reverberant Room Using Harmonic Based Music”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 651-655 (Brighton, UK) (2019 May). https://doi.org/10.1109/ICASSP.2019.8683098.
9. R. Badeau, R. Boyer, and B. David, “EDS Parametric Modeling and Tracking of Audio Signals”, *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx02)*, pp. 139-144 (Hamburg, Germany) (2002 Sep.).
10. Julius O. Smith and Perry R. Cook. 'The Second-Order Digital Waveguide Oscillator', in *Proceedings of the International Computer Music Conference*, pp. 150-153 (San Jose, USA) (1992 Oct.) Revised version http://www.ece.uvic.ca/~bctill/papers/numacoust/Smith_Cook_1992.pdf (accessed May 9, 2021).
11. Bricasti Design, “M7 Stereo Digital Reverberation Processor (rev 5.02.08) Owner’s Manual”. http://www.bricasti.com/images/M7_pdf.zip (accessed May 9, 2021).
12. H. Bass, L. Sutherland, A. Zuckerwar, D. Blackstock, and D. Hester, “Atmospheric Absorption of Sound: Further Developments”, *Journal of the Acoustical Society of America*. vol. 97, no.1, pp. 680-683 (1995 Jan.). https://doi.org/10.1121/1.412989
13. R. Roy and T. Kailath, “Estimation of Signal Parameters via Rotational Invariance Techniques”, *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 7, pp. 984 - 995 (1989 July). https://doi.org/10.1109/29.32276
14. O. Derrien, R. Badeau, and G. Richard, “Parametric Audio Coding With Exponentially Damped Sinusoids”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1489-1501 (2013 July). https://doi.org/10.1109/TASL.2013.2255284
15. R. Badeau, B. David, and G. Richard, "Selecting the modeling order for the ESPRIT high resolution method: an alternative approach", *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. II-1025 - II-1028 (Montreal, Canada) (2004 May). https://doi.org/10.1109/ICASSP.2004.1326435
16. M. Lagrange et al., 'The DESAM Toolbox: Spectral Analysis of Audio', *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx10)*, pp.254-261 (2010 Sep.).
17. M. Abe and J. Smith, “AM/FM rate estimation for time-varying sinusoidal modelling”, *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. iii-201 - iii-204 (Philadelphia, USA) (2005 Mar.). https://doi.org/10.1109/ICASSP.2005.1415681.
18. J. Wells, “Interactive Reverberation Modelling - Supporting Materials” http://www.jezwells.org/publications/reverberation_modelling (accessed May 9, 2021).

19. J. Wells, “Methods for Separation of Amplitude and Frequency Modulation in Fourier-Transformed Signals”, *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx10)*, pp.33-40 (2010 Sep.).
20. X. Serra and J. Smith, “Spectral Modeling Synthesis: A Sound Analysis/Synthesis Based on a Deterministic plus Stochastic Decomposition”, *Computer Music Journal*, vol. 14, no. 4, pp. 12-24 (1990 Winter). https://doi.org/10.2307/3680788
21. S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397-3415 (1993 Dec.). https://doi.org/10.1109/78.258082
22. Z. Liu, J. Li, and P. Stoica, "RELAX-based estimation of damped sinusoidal signal parameters", *Signal Processing*, vol. 62, no.3, pp. 311-321 (1997 Nov.). https://doi.org/10.1016/S0165-1684(97)00132-1
23. V. Vijayan, “Fast SVD and PCA”, https://uk.mathworks.com/matlabcentral/fileexchange/47132-fast-svd-and-pca (accessed May 9, 2021).
24. D. Murphy, and S. Shelley, “OpenAIR: An Interactive Auralization Web Resource and Database”, presented at the *129th Convention of the Audio Engineering Society* (2010 Nov.), Paper 8226.
25. Greg Hopkins. “MTSU Impulse Response Library”, http://bit.ly/hopkinsIR (accessed May 9, 2021).
26. P. Kabal. “An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality”, http://www-mmsp.ece.mcgill.ca/Documents/Reports/2002/KabalR2002v2.pdf (accessed May 9, 2021).
27. J. Wells, “Interactive Reverberation Modelling System”, https://www.jezwells.org/Computer_music_tools.html#revModel (accessed May 9, 2021).
28. Nugen Audio, “Paragon”, https://nugenaudio.com/paragon/ (accessed May 9, 2021).
29. J. Wells, “Interactive Reverberation Modelling System - An overview for users”, https://youtu.be/oTZ6F_2EKVY (accessed May 9, 2021).

7 Author biography

Jeremy (Jez) Wells is an audio designer, artist and engineer. A graduate of the University of Surrey's *Tonmeister* program in Music and Sound Recording, he subsequently gained his PhD, on spectral modelling for creative sound transformation, from the University of York where is currently a Senior Lecturer (Associate Professor) in the Department of Music. He has previously worked for Digital Audio Research and Fairlight, as well as being a Research Associate in acoustics modelling and then a Lecturer in the Department of Electronic Engineering at York. He held an Ingenious Public Engagement Fellowship with the Royal Academy of Engineering from 2011 to 2012.