HIGH-RESOLUTION ANALYSIS AND RESYNTHESIS OF ENVIRONMENTAL IMPACT SOUNDS

L.-M. Reissell Dinesh K. Pai

Computer Science, University of British Columbia and Rutgers University
lmre@cs.rutgers.edu pai@cs.ubc.ca

ABSTRACT

Impact sounds produced by everyday objects are an important source of information about contact interactions in virtual environments and auditory displays. Impact signals also provide a rich class of real and synthetic percussive musical sounds. However, their perceptually acceptable resynthesis and modification requires accurate estimation of mode parameters, which has proved difficult using traditional methods.

In this paper we describe some of the problems posed by impact phenomena when applying standard methods, and present a phase-constrained high-resolution algorithm which allows more accurate estimation of modes and amplitudes for impact signals. The phase-constrained algorithm is based on least squares estimation, with initial estimates obtained from a modified ESPRIT algorithm, and it produces better resynthesis results than previously used methods. We give examples with everyday object impact sounds.

[Keywords: sound, estimation, synthesis, contact, impact, conditioning, ESPRIT, least squares]

1. INTRODUCTION

Impact sounds are ubiquitous and carry very useful information that can be used for recognition of the structure and material properties of the environment, and of the nature of the impact [15, 22]. They are critical in human perception of contact [10, 13, 17]. Impact sounds are also increasingly important in music [5], computer graphics [27, 18, 12], and multimodal user interfaces [4, 28]. In haptics, impact sounds can be used to complement the force feedback of a haptic device [8, 9], and impact vibrations have been used to improve performance [14, 19, 16].

The impact signal produced by an object is characterized by a discrete set of decaying vibration modes which depend on material properties and boundary conditions. The signal can be modeled as the real or imaginary part of a sum of complex decaying exponentials $z^t = e^{(a+iw)t}$, or *modes*, determined by complex *poles* $z = e^{a+iw}$, with complex amplitude coefficients. The problem of estimating exponential modes from noisy data is well studied in signal processing and system identification, and a number of successful high resolution algorithms — algorithms which can resolve frequencies beyond the discrete Fourier transform (DFT) resolution limit — have been developed for this and related tasks.

However, impact signals pose difficulties in the application of these methods. Unlike the usual contexts in which these methods have been applied, impact signals are brief and possess fast-decaying modes. Another difficulty is closely spaced modes, which are common due to near-symmetries in the vibrating objects. These characteristics lead to serious conditioning problems and to large inaccuracies in mode analysis from typically available data lengths, even when no other noise is present. Using a lower-rank approximation (fewer modes) to correct conditioning problems leads to mode errors, which are insignificant for the short data, but can audibly distort longer resynthesis.

The artificial "impact sound" in Fig. 1 illustrates some of these problems. The short test signal can be modeled well by highly incorrect modes chosen at random; these random modes also contain non-decaying modes. However, the exponentially growing modes sound nonsensical in longer synthesis.

In this paper we discuss the accurate off-line estimation of impact signal modes when only short data windows are available for analysis. We consider the conditioning properties of impact signals, and their implications for the choice of analysis method. We observe that, for many impact signals, the mode-amplitude determination problem is nearly ill posed.

To allow more accurate mode analysis of difficult impact sounds from short data, we propose an improved high-resolution method. The method is based on well-known existing methods, least squares estimation and ESPRIT, modified for impact sounds, and it models impact sounds better. Experimental results are presented in this paper. They can be seen and heard at a supplemental web page [26].

2. RELATED WORK

The general problem of estimating the spectral content of a noisy signal has been extensively studied; see, for instance, [25] for a recent overview. The obvious method is to estimate the spectra by picking the peaks of the discrete Fourier transform (DFT) of a windowed signal, but the frequency resolution of such an approach is limited. This problem has been addressed with the development of so-called *high resolution* methods, whose accuracy transcends the DFT resolution limit, and which can be used on relatively short data samples.

Several different classes of methods have been developed for high resolution estimation and related problems in signal processing and system identification [25], [7].

Common and effective methods for estimating signal parameters include ESPRIT, MUSIC, and Least Squares (LS) and its generalization, Maximum Likelihood (ML). In some settings these can be formulated as similar optimization problems [7]. LS is an important instance of methods which attempt to directly minimize the distance between a low rank shift-invariant subspace and the observed signal. ESPRIT and MUSIC are instances of *subspace methods*, which rely on the eigendecomposition of the data covariance matrix to determine an estimate for the signal space. These latter methods do not require non-linear optimization to implement. Most approaches separate the problems of mode estimation and amplitude estimation. (Amplitude estimation from known

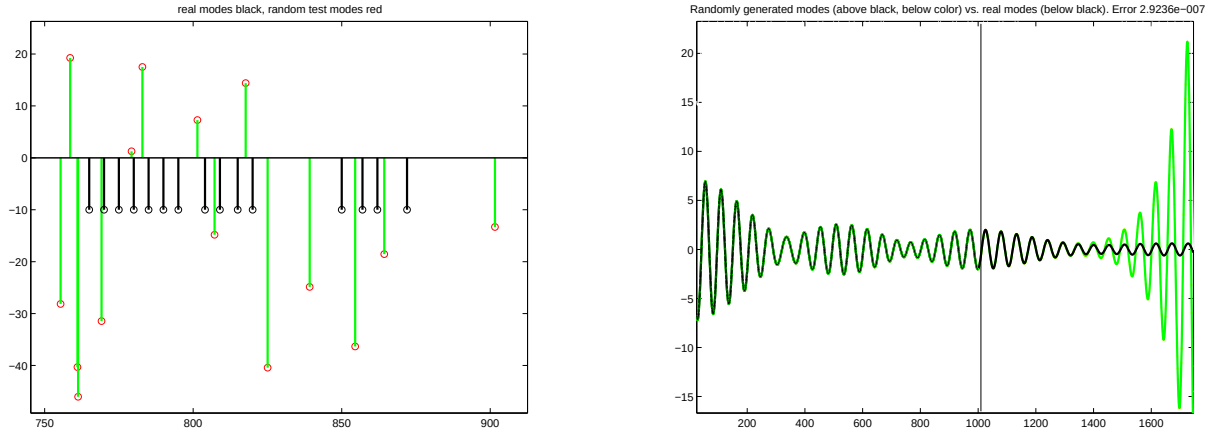


Figure 1: *Randomly picked modes with a wide mode error can approximate test data well. Modes are depicted in a stem plot, frequency on x-axis, modulation exponent ("decay") on y-axis. Decaying modes point in the negative y-direction. (a) Original modes (black) and randomly generated modes (green), some of which do not decay. (b) Signals synthesized from original modes (in black) and from random modes (in color). Signal approximation from random modes is excellent for the test signal (samples to the left of the vertical line), but gives nonsensical results in longer synthesis. Test data approximation error is 2.9e-007.*

modes for non-decaying sinusoidal signals has been discussed in detail in [24].)

All the above estimation methods are unbiased and have good asymptotic properties, especially in the presence of Gaussian noise and when the modes do not decay too sharply. For recent results on ESPRIT, the most statistically accurate of the subspace methods for the harmonic retrieval problem, see [1], [3].

The methods present tradeoffs in estimation accuracy and computational complexity. ESPRIT methods are widely used due to their relatively efficient closed-form solution and good asymptotical properties [25]. However, in the presence of nearby modes, ESPRIT is poorly conditioned and requires very long data samples for accuracy, reducing the high-resolution benefits of the method. In addition, the computational complexity of ESPRIT increases as the $O(N^3)$ required for the usual SVD calculation (although some improvements can be made). Accordingly, to reduce model and data size, sound samples are often preprocessed for ESPRIT by filtering. Also, for musical samples, perceptually important higher frequency modes can be enhanced at the expense of lower frequency "noise." However, impact sounds, which have denser spectra, can lose important timbre information in such preprocessing.

As opposed to ESPRIT, Least Squares (LS) estimation, the Gaussian noise form of Maximum Likelihood estimation (ML), requires solving a nonlinear optimization problem. This is generally computationally expensive. The method requires good global search algorithms and initial estimates to converge correctly[25]. However, in the presence of closeby poles, the method is more accurate than ESPRIT. The applicability of the standard version of LS for impact sounds is reduced by conditioning problems and spurious minima.

In previous work on estimating models of impact sounds [22, 20], mode frequencies were first estimated from power spectra and peak identification. The modes could be pruned based on perceptual criteria [29]. The results provide a convincing method for modeling moving impact locations, but the perceptual accuracy of the synthesized timbres can also be improved. [6] presents a fast but approximative method, which fits a small number of modes to the signal near each chosen frequency peak. The method is designed to be approximative and it is not clear how it will perform for dense modes.

We applied an earlier version of our algorithm to haptic force signals in [21]. In the current paper, the algorithm has been extended to deal with difficult sound examples, using shorter signal samples and more complicated spectra. We give more details of the conditioning problems. Added examples and the new results section give insight into the behavior of the algorithms and into the reasons for looking for improvements to existing methods.

3. IMPACT SIGNAL MODELS

We model the measured impact signal y as the real or imaginary part of a complex model signal x, together with additive noise ν:

$$y(t) = \mathrm{Re}(x(t)) + \nu \quad \text{or} \quad y(t) = \mathrm{Im}(x(t)) + \nu, \qquad (1)$$

where we assume that the noise is Gaussian with variance σ^2.

The model signal x is the sum of K decaying exponential signals produced by the vibration *modes* of the sound generating objects,

$$x(t) = \sum_{k=1}^{K} c_k \; e^{(a_k + i\omega_k)t} = \sum_{k=1}^{K} c_k z_k^t. \qquad (2)$$

Here, c_k is the complex amplitude of mode k, consisting of both magnitude and phase; $z_k = e^{a_k + i\omega_k}$ is a complex pole of the z-transform of x, a_k is the decay parameter and ω_k the frequency. The poles are assumed distinct here. Note that the number of modes K, the *model order* of the problem, is generally unknown and must be estimated.

The *signal subspace* consists of the space spanned by the K modes $\boldsymbol{b}_k(t) = z_k^t$. In discrete time, each mode can be written, for a window of M samples, as the vector $\boldsymbol{b}_k = ({z_k}^0, {z_k}^1, {z_k}^2, ..., {z_k}^{M-1})^T$. Then, the signal subspace is the column space of the Vandermonde matrix

$$\boldsymbol{B} = (\boldsymbol{b}_1 \; \boldsymbol{b}_2 \; ... \; \boldsymbol{b}_K). \qquad (3)$$

Note that $\boldsymbol{B}$ is a function of the poles, $\boldsymbol{z} = (z_1, \ldots, z_K)^T$; we will write $\boldsymbol{B}(\boldsymbol{z})$ when it is important to highlight this dependence.

When highlighting the dependence of $\boldsymbol{B}$ on column length M, we will use $\boldsymbol{B}_M$.

Similarly, we can define the sampled signal $\boldsymbol{x}(t) = (x(t)\ x(t+1)\ \ldots\ x(t+M-1))^T$; $\boldsymbol{y}(t)$ is defined similarly. This notation allows the signal model in Eq. 2 to be expressed compactly as

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{c}, \tag{4}$$

where $\boldsymbol{c}$ is the vector of complex mode amplitudes $(c_1, \ldots, c_K)^T$.

Existing performance analyses for the line spectrum analysis problem indicate that in the presence of Gaussian noise, two of the most statistically accurate estimation methods are least squares and ESPRIT.

In nonlinear Least Squares (LS) estimation, which is equivalent to Maximum Likelihood estimation (ML) for Gaussian noise, the problem is to estimate the signal parameters $\boldsymbol{z}$ and $\boldsymbol{c}$ so that the error between the data $\boldsymbol{y}$ and its projection onto the signal space $\boldsymbol{B}$ is minimized:

$$\boldsymbol{z}, \boldsymbol{c} = \arg\min \parallel \boldsymbol{y} - \boldsymbol{B}(\boldsymbol{z})\boldsymbol{c} \parallel^2 . \tag{5}$$

This problem is separable into two optimization problems, one for modes and a linear one for the amplitudes. This is because, for any set of poles $\boldsymbol{z}$ the optimal choice in the least squares sense for the coefficients $\boldsymbol{c}$ is found via the pseudoinverse $\boldsymbol{B}^\dagger$ of $\boldsymbol{B}$:

$$\boldsymbol{c}(\boldsymbol{z}) = \boldsymbol{B}(\boldsymbol{z})^\dagger \boldsymbol{y}, \tag{6}$$

where $\boldsymbol{B}^\dagger = (\boldsymbol{B}^*\boldsymbol{B})^{-1}\boldsymbol{B}^*$, where '*' indicates the conjugate transpose. This is an unbiased linear estimate with minimal variance for Gaussian noise [24], and can be used with any method which provides an estimate of the modes.

For LS, the mode estimation can be performed by substituting $\boldsymbol{c}(\boldsymbol{z})$ into the minimization (Eq. 5), resulting in an amplitude-independent formulation of the problem:

$$\begin{aligned} \boldsymbol{z} &= \arg\min \parallel (\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^\dagger)\boldsymbol{y} \parallel^2 = \\ &= \arg\max_{\boldsymbol{z}} (\boldsymbol{y}^* \boldsymbol{B}\boldsymbol{B}^\dagger \boldsymbol{y}). \end{aligned} \tag{7}$$

The amplitude-dependent formulation minimizes the error between the observed signal $\boldsymbol{y}$ and its projection to space $\boldsymbol{B}$; the amplitude-independent formulation, equivalently, maximizes the alignment between the signal and its projection.

In mode estimation, an alternative to dealing with the sensitive optimization of Eq. 7 is to use shift invariance of the underlying signal modes and the covariance matrix of the data, defined as $\boldsymbol{R} = E\{\boldsymbol{y}\boldsymbol{y}^*\}$. The estimated covariance matrix is a square matrix, whose dimension is denoted here by $M \times M$. The eigenstructure of $\boldsymbol{R}$ contains all the information needed to estimate $\boldsymbol{z}$; specifically, the eigenvectors corresponding to the K dominant eigenvalues of $\boldsymbol{R}$ span the signal subspace of $\boldsymbol{B}$. ESPRIT, for example, provides a method for estimating $\boldsymbol{z}$ this way. We sketch the main steps of the algorithm below.

Using the eigendecomposition of the covariance matrix is equivalent to using the singular value decomposition (SVD) of the matrix $\boldsymbol{Y}$, formed from successive windowed observations $\boldsymbol{y}_t$ of the data $\boldsymbol{y}$, $\boldsymbol{y}_t = (y(t)y(t+1)\ldots)^T$. $\boldsymbol{Y}$consists of W successive shifted data windows, starting at a given time t:

$$\boldsymbol{Y} = (\boldsymbol{y}_t\ \boldsymbol{y}_{t+1}\ \cdots\ \boldsymbol{y}_{t+W-1}). \tag{8}$$

With $\boldsymbol{y}$ window length M, the covariance matrix $\boldsymbol{R}$ can be estimated as the $M \times M$ matrix $\frac{1}{W}\ \boldsymbol{Y}\boldsymbol{Y}^*$. The choice of window shift direction leads to two similar formulations for $\boldsymbol{Y}$, where the matrix $\boldsymbol{Y}$ has Toeplitz or Hankel structure. (For more details, in the Hankel formulation, see e.g. [1].)

4. ESPRIT ESTIMATION OF SIGNAL MODES

Our mode estimation algorithm is an optimization procedure tailored for impact signals. The initial modes for the optimization are obtained from an algorithm which modifies the results from standard ESPRIT [23] to account for the properties of impact signals. (Other methods for initial mode determination can be used, but ESPRIT has the advantage of being asymptotically accurate enough so that the optimization step may not be necessary for longer data samples and sparser spectra.)

Standard ESPRIT, which we describe briefly below, is an easily implemented algorithm based on the eigendecomposition of the data covariance matrix, or equivalently, on the SVD of the Toeplitz or Hankel matrix $\boldsymbol{Y}$. The method uses the shift invariance of the signal subspace and does not require nonlinear optimization. For more details, see e.g. [25].

In ESPRIT, the eigenvectors $\boldsymbol{U}$ of the covariance matrix $\boldsymbol{R}$ are separated into two subspaces, an estimate $\boldsymbol{S}$ of the signal subspace $\boldsymbol{B}$, and the noise subspace $\boldsymbol{N}$. The signal subspace is determined by the model order, or signal subspace rank, K. If the noise ν is gaussian, of variance σ^2, the covariance matrix of the observed signal $\boldsymbol{y}(t)$ is the sum of the covariance matrices of $\boldsymbol{B}\boldsymbol{c}$ and the noise covariance matrix $\sigma^2\boldsymbol{I}$. From this it follows that the eigenvalues of $\boldsymbol{R}$ corresponding to the signal subspace of rank K are the K largest eigenvalues.

Finding the poles from $\boldsymbol{S}$ relies on the shift-invariance of the exponential mode. Shift-invariance of the signal space $\boldsymbol{B}$ is expressed as

$$\boldsymbol{B}^{FRD} = \boldsymbol{B}^{LRD} diag(\boldsymbol{z}), \tag{9}$$

where $\boldsymbol{B}^{FRD}$ and $\boldsymbol{B}^{LRD}$ are $\boldsymbol{B}$ with the first and last row deleted, respectively. Since the corresponding eigenvectors $\boldsymbol{S}$ also span the signal subspace $\boldsymbol{B}$, $\boldsymbol{S}$ is also shift-invariant, and we have $\boldsymbol{S} = \boldsymbol{G}\boldsymbol{B}\boldsymbol{G}^{-1}$ for a basis-change matrix $\boldsymbol{G}$. This basis change takes the poles $diag(\boldsymbol{z})$ of the original space $\boldsymbol{B}$ to a matrix $\Phi = \boldsymbol{G}diag(\boldsymbol{z})\boldsymbol{G}^{-1}$ with the poles as eigenvalues, and the shift-invariance of Eq.9 can be expressed as:

$$\boldsymbol{S}^{FRD} = \boldsymbol{S}^{LRD}\Phi \tag{10}$$

The poles $z_k, k = 1, \ldots, K$ are the eigenvalues of the matrix Φ, found as a least squares or total least squares solution.[25]

The complex amplitudes $\boldsymbol{c}$ are determined from the data $\boldsymbol{y}$ and the estimated signal subspace matrix $\boldsymbol{B}$ using the separability criterion of Eq. 6.

For accurate estimation, model order K must be chosen to be equal to (or larger than) the actual dimension of the signal space; with a lower model order, the estimation results are approximations. There are several methods for determining model order [25], [2].

The computational complexity of the ESPRIT algorithm is bounded by the cost of the eigenvalue computation for the $M \times M$ covariance matrix, or equivalently, the SVD computation for $\boldsymbol{Y}$ of Eq. 8, $O(M^3)$ in standard forms. Computationally it is frequently better to obtain eigenvectors of the covariance matrix $\boldsymbol{R}$ directly from the SVD of the matrix $\boldsymbol{Y}$.

5. CONDITIONING PROBLEMS IN IMPACT SIGNAL ESTIMATION

Impact signals pose difficulties in the application of standard spectrum analysis methods like ESPRIT and ML/LS. Impact signals are characterized by spectra that are dense or, due to near symmetries in the objects, locally dense, and with modes that decay fast or at widely different rates. These characteristics cause conditioning problems which make accurate mode determination very difficult in the presence of noise or truncation errors.

We summarize some of the algorithmic problems in impact sound mode estimation below.

- Basis matrix conditioning: for many observed impact signals, the mode-amplitude determination problem of Eq. 4 is nearly ill posed.

  The conditioning of the Vandermonde basis matrix $\boldsymbol{B}$ of Eq. 3 becomes worse when the poles approach each other, when the modes decay more sharply, and when column length (sample length) decreases. When estimating impact signals from short samples, all of these factors apply. The original signal can then be approximated within a small error also from nearby, incorrect mode spaces.

  One solution is to reduce the number of modes used in the approximation. But, although the original short samples can then be adequately represented by such low rank approximations, or from incorrect modes in general, the induced mode errors may distort new synthesized examples. This happens for example when generating longer sounds (Fig. 1) or when different subsets of the signal modes are activated during impact, or when the sound is modified in applications such as pitch-shifting.

  So mode errors can be perceptually relevant, even if the test data error norm is small.

- Covariance matrix conditioning

  The conditioning of the covariance matrix $\boldsymbol{R}$ and the conditioning of the matrix $\boldsymbol{Y}$ of Eq. 8 depend on the conditioning of $\boldsymbol{B}$ and the complex amplitudes of the signal. This can be seen from the following relationship between the matrices $\boldsymbol{B}$ and $\boldsymbol{Y}$. In this discussion, we assume that $\boldsymbol{Y}$ is formed from the complex noiseless signal.

  If we assume that $\boldsymbol{Y}$ is formed from W-many data windows of length M, taken from time index t onwards, where t is indexed as $t = 0, 1, 2, \ldots$, we can write

  $$\boldsymbol{Y} = \boldsymbol{B}_M diag(\boldsymbol{z}^t) diag(\boldsymbol{c}) \boldsymbol{B}_W^T, \qquad (11)$$

  where $\boldsymbol{c}$ are the complex amplitudes of the signal, and $\boldsymbol{B}_M$, $\boldsymbol{B}_W$ are Vandermonde basis matrices with column lengths M, W (see Sec. 3).

  From this formulation it can be seen that the maximum rank of $\boldsymbol{Y}$ is the maximum rank K of $\boldsymbol{B}$, and that (a sufficiently large) $\boldsymbol{Y}$ has full rank exactly when the poles $\boldsymbol{z}$ are distinct and the amplitudes $\boldsymbol{c}$ are non-zero. The same result is seen to hold for $\boldsymbol{R}$ by writing out the covariance matrix expression for $\boldsymbol{R}$ in terms of $\boldsymbol{Y}$. (See also [11, 1].)

- ESPRIT conditioning

  Given a fixed data sample length, ESPRIT will be inaccurate when the covariance matrix $\boldsymbol{R}$ is poorly conditioned; that is, when the mode basis is poorly conditioned or the mode amplitudes are small.

  In this case, small errors due to data truncation or noise can make the determination of the eigenvectors $\boldsymbol{U}$and of the signal subspace, estimated as the span of a subset of $\boldsymbol{U}$, widely inaccurate. Algorithms for determining model order using the covariance matrix eigendecomposition ([1]) can also miss.

  In poorly conditioned ESPRIT a portion of the signal subspace determined by ESPRIT is noise.

  An example is given in Figs. 2, 3. In Fig. 2(b) the "nonsense" modes correspond to noisy eigenvectors in $\boldsymbol{U}$. Fig. 3(b) shows the consequence of poor ESPRIT conditioning on signal approximation: the test signal has a good low-rank approximation from the decaying ESPRIT modes, but decay and pitch errors are visible and audible in longer resynthesis.

- Optimization

  For the same reasons as above, closeby modes and sharp decays yield "too many near solutions" to the general optimization problem. As the Vandermonde matrix becomes more rank-deficient, these near-solutions become more accurate, and, similarly, spurious local error minima approach zero. It is very difficult to get convergence to the correct modes in optimization algorithms unless the problem is regularized further.

The difficult conditioning properties of impact signals have some implications in the choice of analysis method when accurate mode modeling is important:

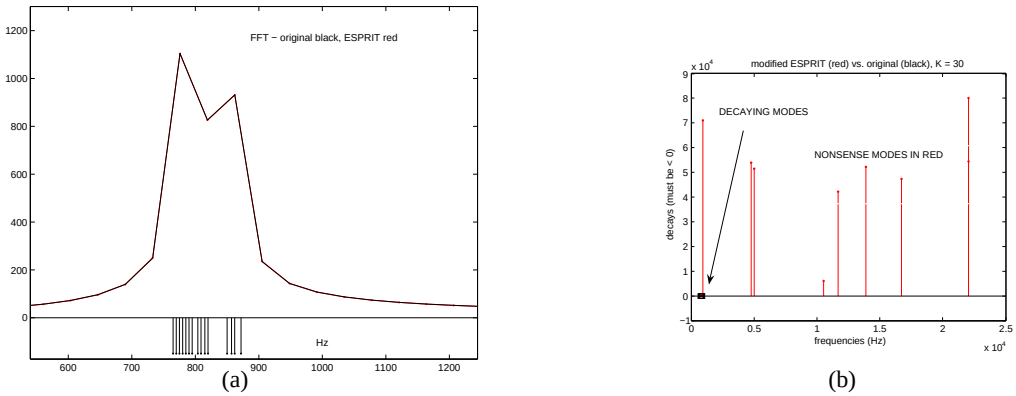
- For impact signals, perceptually acceptable mode accuracy is not guaranteed by a small 2-norm error for the test signal.
- For dense spectra and sharp decays, cross-validation or further optimization, possibly from multiple initial values, is needed to ensure that a mode value near the global optimum has been found.
- One strategy is to use a large number of modes with further constraints to regularize the optimization problem and to reduce the number of spurious minima. We use a phase-constraining strategy below. Another strategy is low-rank approximation. Low-rank mode approximations can be very useful when the decays for dense modes are so sharp that the low-rank estimation is accurate even with larger sample lengths. However, ensuring the perceptual correctness of a low-rank approximation usually requires cross-validation or careful psychoacoustic analysis.

6. PHASE-CONSTRAINED MODE ESTIMATION FOR IMPACT SIGNALS

We address the difficulties described above by using phase constraints.

We constrain the LS/ML algorithm with a physically based constraint and require that the phases of the modes equal zero or π. This effectively aligns the signals with the time of impact, and restricts the amplitude to real numbers. This exploits an important piece of prior information available about impact signals not present in many other types of signals, and reduces the number of local minima of the target function.

Due to the reduced number of local minima, phase-constrained optimization is not as sensitive to initial value placement as optimization allowing arbitrary phase, but the LS algorithm still requires relatively good initial values to converge correctly.



(a) (b)

Figure 2: *Problems with DFT peak finding and ESPRIT. (a) FFT magnitude of test sound (only 2 maxima visible for 15 modes, shown in a frequency-decay stem plot). (b) The results of the unmodified ESPRIT method. Only the "nonsense" modes unrelated to signal content are visible - the signal modes and the 5 decaying ESPRIT modes occupy the small black box in the image.*

We use ESPRIT to produce the initial values, modified to allow its use for impact signals despite poor conditioning. (For general signals, this modification is not possible.) The essential step is the obvious one of removing non-decaying modes from the results of the ESPRIT algorithm. In addition, we assure that the conditioning of the resulting Vandermonde matrix $\boldsymbol{B}$ is better than a given tolerance both by removing modes and by normalization.

Since with impact signals ESPRIT usually reduces problem rank and can miss significant modes, we supplement the initial mode estimates from ESPRIT by frequencies from DFT peak finding. (One should note that the best low-rank approximations usually do not satisfy the same physical constraints as the signal modes: for example, the ESPRIT rank-deficient approximation modes do not generally have zero phase, even if the original modes do.) The advantage of using ESPRIT is its asymptotic accuracy for most data.

Finally, we refine the modes by solving the nonlinear optimization problem of Eq. 5, with the phase constraint.

Modified ESPRIT alone can be sufficient for impact signal modeling, especially if a large sample is available for the estimation. But for difficult, fast-decaying impact signals with closeby modes, this phase-constrained LS optimization with modified ESPRIT initial values produces better resynthesis results than previously used methods.

6.1. A summary of the algorithm:

For phase-constrained optimization, the signal is modeled as the imaginary part of the complex signal.

A. Initial Mode Determination from Modified ESPRIT

1. Determine a model order K, and the covariance matrix size, M.

   Since the signal is real, and the model is complex, the algorithm will use $K = 2K^*$, where K^* is the number of real exponentials expected in the data. We let $M = (N+1)/3$.

2. Determine the estimates for the poles $z_k, k = 1, \ldots, K$ as in standard ESPRIT.

3. Form a subset of the modes z_k of length K', by discarding all nondecaying modes; renumber the modes $z_k, k = 1, \ldots, K'$.

4. Determine the estimated Vandermonde matrix $\boldsymbol{B}$ for the remaining modes.

5. Determine the amplitudes and phases from the complex amplitudes $\boldsymbol{c}$, the data $\boldsymbol{y}$, and the estimated signal subspace matrix $\boldsymbol{B}$ using the optimization separability criterion: $\boldsymbol{c} = \boldsymbol{B}^{\dagger}\boldsymbol{y}$.

6. Retain only the $K'' = K'/2$ complex conjugate poles z_k with positive imaginary part, and their coefficients c_k. Renumber the modes as $z_k, k = 1, \ldots, K''$.

7. Given c_k and the poles z_k, $k = 1 \ldots K''$, the estimated signal model is then

$$\boldsymbol{x}(t) = \sum_{k=1}^{K''} \alpha_k e^{\phi_k i}\, e^{(a_k + i\omega_k)t} \tag{12}$$

   where the decays and frequencies a_k, ω_k are obtained from the poles z_k, and the amplitudes and phases from the complex coefficients c_k.

B. Add missed DFT peaks

If DFT peak frequencies are missed to a given tolerance by step A, add modes corresponding to these peak frequencies.

(There are several methods for approximating the corresponding mode decays, but for initial modes, using an average mode decay has been sufficient.)

C. Zero phases and determine initial real amplitudes

Zero all phases, and use the real form of the separability condition Eq. 6 to determine the best (real) amplitudes for the initial modes.

D. Phase Constrained Least Squares

Using modes obtained in the previous step as initial values, solve the nonlinear optimization problem of Eq. 5, with the phase

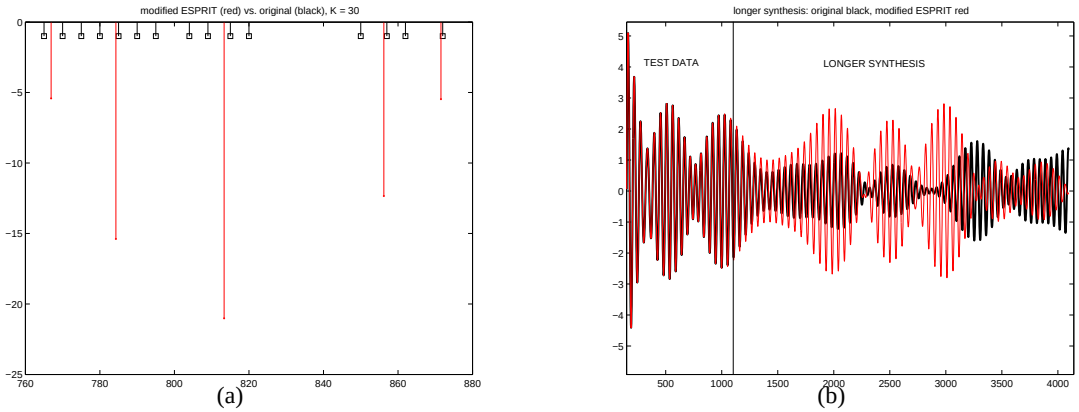


Figure 3: *Modified ESPRIT. (a) Original modes (black) and the 5 modified ESPRIT modes (red). (b) In the modified ESPRIT algorithm, pruned modes are used to produce a good fit to the test signal (to the left of vertical line). Mode errors become apparent with a longer synthesis; here ESPRIT (red) diverges from the correct signal (black). Pitch and decay errors are clearly heard in longer synthesis. Our optimization algorithm addresses this problem.*

constraint. For single impact sounds, the phase constraint is implemented simply by allowing the amplitudes c_k in the signal model Eq. 2 to take any real values, which constrains the phase variable to zero or π.

We used the Matlab function `lsqnonlin`.

To allow for better performance and to reduce the number of variables in optimization, it is in many cases possible to carry out the bulk of the optimization in band-limited steps, finetuning the results in a final global step. (Straightforward band-limited optimization will obviously be the more accurate the better the modes separate into mutually nearly orthogonal signal subspaces.)

7. RESULTS

For the examples below, the sounds and additional data are also available at a supplemental webpage [26].

7.1. Artificial impact sounds

We illustrate ESPRIT conditioning problems with an example of an artificial "impact" sound consisting of 15 modes between 765 and 872 Hz in frequency. The modes form two groups which are visible from the DFT, Fig.2(a), but the modes are sufficiently closely spaced so that individual modes cannot be picked out.

To illustrate the effects of conditioning with truncated data, we do not add noise. The data length here is $N = 1024$ for clearer illustration, but the behavior of ESPRIT is analogous for longer data lengths. The covariance matrix size is $M = (N + 1)/3$ (for reasons for the choice, see [1]).

For this test signal, ESPRIT will be inaccurate for all parameter choices, because the eigenproblem of the covariance matrix is poorly conditioned and sampled matrices have truncation errors. ESPRIT misses decaying modes. In Fig.2 and Fig.3, 10 of the 15 original modes estimated by ESPRIT are "nonsense" modes, or non-decaying modes clearly outside the frequency range. Here modified ESPRIT provides a low-rank approximation to the data, Fig.3(a, b). However, mode errors distort longer synthesis, Fig.3(b).

7.2. Real impact sound

Our algorithm performs well even in the more difficult case of a real impact sound. Since it is difficult to convey the accuracy of results of sound synthesis with figures, we have collected the sounds and other figures that could not be included due to space limitation on the website.

We recorded sounds from everyday objects by tapping them in different locations. The objects include wine glasses, a four-sided plastic office trashcan tapped with a flick of a finger, and a plastic water bottle struck with a wooden striker.

The recording was performed in an acoustic isolation chamber (built by Eckel Noise Control Technologies, Cambridge, MA), using a 1/2 inch condenser microphone type 4189 (Brüel and Kjær, Denmark). The signal was digitized at 44.1 KHz, 16 bit resolution, using an NI DAQCard-6036 (National Instruments, Austin, TX) and Matlab's Data Acquisition Toolbox (The MathWorks, Natick, MA). Microphone placement varies for the different objects, for example, for the bottle example, it was held approximately 50cm from the middle of the bottle. The data lengths used for estimation are standard, varying from 37 ms to 68 ms. The sound was not preprocessed for the algorithms.

7.2.1. Wine glasses, a bottle, and a trashcan

We resynthesize sounds from tapping on a plastic trashcan, wineglasses, and a plastic bottle. The different modal content of the sounds is reflected in the difficulty of analysis.

For the wineglass sounds, which have sparse spectra, Fig. 4(a), modified ESPRIT alone produces very good results. The ESPRIT result is perceptually difficult to distinguish from the original (both on paper and in sound). For the other sounds, phase-constrained optimization produces audibly the best results. For the bottle sound, where we have used a long data sample for analysis, the modified ESPRIT version is perceptually quite similar to the original, but mode inaccuracies are still heard in longer resynthesis.

The trashcan example has the densest mode spectrum, Fig. 4(b), and phase-constrained optimization is clearly better than ESPRIT here. The resynthesis results are depicted in Fig. 4(c). The estimated modes were used to resynthesize longer data. In this cross-

validation, the phase-constrained optimization result is perceptually difficult to distinguish from the original. Phase-constrained optimization converges even when initial values are determined from ESPRIT with a relatively arbitrarily chosen model order.

7.3. Synthetic Sound Modification Example

Accurate mode analysis can be used also as a starting point for compression, or to synthesize artificial but realistic impact sounds which modify the mode structure of existing sounds. The modifications can include, for example, pruning modes selectively or changing the behavior of the modes of one sound to mimic specified features of another.

As a simple example, we compress sound representation by starting with the full set of estimated modes and removing all modes with decays faster than a preset decay limit. Removing sharply decaying modes also softens the attack. For sounds like the wineglass sound, the results are perceptually close to the original with only a few modes retained (12 out of 83). By changing the decays of the slowest-decaying modes only, we can also change the decay properties of the sound without affecting attack quality. The modified sounds, which are realistic, can be heard on the website.

8. CONCLUSIONS

Accurate estimation of the parameters of impact sounds is very difficult due to the poor conditioning of the problem, and traditional approaches have not produced perceptually accurate resynthesis results from short samples. We have outlined these conditioning problems, and described a high resolution method based on ESPRIT and least squares estimation, with phase constraints. The method gives excellent resynthesis results even with difficult impact signals and short samples. Example sounds are available at [26].

9. ACKNOWLEDGMENTS

This work was supported in part by NIH grant R01NS050942, NSF grants IIS-0308157, ACI-0205671, and EIA-0321057, and the Canada Research Chairs program.

10. REFERENCES

[1] R. Badeau. *Méthodes à haute résolution pour l'estimation et suivi de sinusoïdes modulées. Application aux signaux de musique.* Ph.D. Thesis, l'Ecole Nationale Supérieure des Télécommunications, 2005.

[2] R. Badeau, B. David, and G. Richard. High resolution spectral analysis of mixtures of complex exponentials modulated by polynomials. *IEEE Trans. Signal Processing*, 54(4):1341–1350, April 2006.

[3] R. Badeau, B. David, and G. Richard. A new perturbation analysis for signal enumeration in rotational invariance techniques. *IEEE Trans. Signal Processing*, 54(2):450–458, February 2006.

[4] W. Buxton. Using our ears: an introduction to the use of nonspeech audio cues. In E. Farrell, editor, *Extracting meaning from complex data: processing, display, interaction. Proceedings of the SPIE*, volume Vol 1259, pages 124–127, 1990.

[5] P. Cook. *Real Sound Synthesis for Interactive Applications*. A.K. Peters, 2002.

[6] R. Corbett, K. van den Doel, J. E. Lloyd, and W. Heidrich. Timbrefields — 3d interactive sound models for real-time audio. http://www.cs.ubc.ca/∼kvdoel/publications/preprint2.pdf, 2006.

[7] A.-J. V. der Veen, E. F. Deprettere, and A. L. Swindlehurst. Subspace based signal analysis using singular value decomposition. *Proc. of IEEE*, 81(9):1277–1308, September 1993.

[8] D. DiFilippo and D. K. Pai. The ahi: An audio and haptic interface for contact interactions. In *UIST'00 (13th Annual ACM Symposium on User Interface Software and Technology)*, November 2000.

[9] D. E. DiFranco, G. L. Beauregard, and M. A. Srinivasan. The effect of auditory cues on the haptic perception of stiffness in virtual environments. In *Sixth Annual Symposium on Haptic Interfaces for Virtual Environments and Teleoperator Systems*, pages DSC–Vol. 61, pp. 17–22. ASME, 1997.

[10] W. W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993.

[11] Y. Hua and T. Sarkar. Matrix pencil methods for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. Acoust. Speech Signal Processing*, 38(5):814–824, May 1990.

[12] D. L. James, J. Barbic, and D. K. Pai. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Transactions on Graphics*, 25(3):987–995, July 2006.

[13] R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Hearing material: Perception of material from contact sounds. *PRESENCE: Teleoperators and Virtual Environments*, 9(4):399–410, August 2000.

[14] D. A. Kontarinis and R. D. Howe. Tactile display of high-frequency information in teleoperation and virtual environments. *Presence*, 4(4):387–402, 1995.

[15] E. Krotkov and R. Klatzky. Robotic perception of material: Experiments with shape-invariant acoustic measures of material type. In *Preprints of the Fourth International Symposium on Experimental Robotics, ISER '95*, Stanford, California, 1995.

[16] K. J. Kuchenbecker, J. Fiene, and G. Niemeyer. Improving contact realism through event-based haptic feedback. *IEEE Trans. Vis. Comput. Graph*, 12(2):219–230, 2006.

[17] S. Lederman and R. Klatzky. Multisensory texture perception. In G. Calvert, C. Spence, and B. Stein. Handbook of Multisensory Processes. MIT Press: Cambridge., 2004.

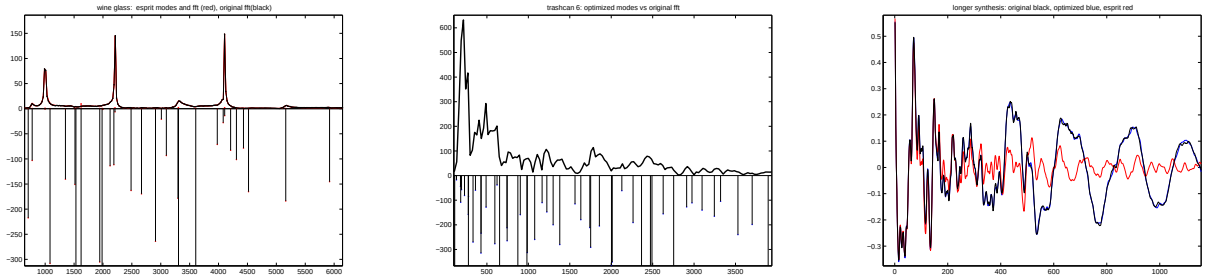
[18] J. F. O'Brien, P. R. Cook, and G. Essl. Synthesizing Sounds From Physically Based Motion. In *Proceedings of ACM SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 529–536, Aug. 2001.

[19] A. Okamura, J. Dennerlein, and M. Cutkosky. Reality-based models for vibration feedback in virtual environments. *ASME/IEEE Transactions on Mechatronics, Focused Section on Haptic Devices and Applications*, 6(3):245–252, 2001.

[20] D. K. Pai et al. Scanning physical interaction behavior of 3D objects. In *SIGGRAPH*, pages 87–96, 2001.

[21] L.-M. Reissell and D. K. Pai. High resolution analysis of impact sounds and forces. In *WHC'07 (World Haptics Conference)*, March 2007.

[22] J. L. Richmond and D. K. Pai. Active measurement and modeling of contact sounds. In *Proceedings of the 2000 IEEE International Conference on Robotics and Automation*, pages 2146–2152, San Francisco, April 2000.



(a) (b) (c)

Figure 4: *(a) Tapping on a wineglass, data length N = 3000. FFT of original (black) and modified ESPRIT (red) shown superimposed, original on top, modified ESPRIT (red) modes on bottom. The zoom shows the two FFTs to be close to identical. Estimated modes are plotted in stem plot (red) at frequency locations, with decay given by the length of the stem. (b) Tapping on a plastic trashcan, data length N = 1630. FFT of original signal, with a stem plot of the optimized modes (black). Modes are densely spaced also beyond the zoomed view. (c) Tapping on a plastic trashcan, data length N = 1630, zoom. Original data (black), phase-constrained optimization (blue), modified ESPRIT (red). Longer data has been resynthesized from estimated modes and compared in cross-validation with unused portions of the original sound. Test data is to the left of the vertical line. The modified ESPRIT algorithm produces slightly incorrect pitch and decay; phase-constrained optimization is nearly identical to the original. (The optimization procedure has been stopped before full convergence and the ESPRIT model order in the data in Fig. 4 has been chosen relatively arbitrarily; the results are representative but not optimal.)*

[23] R. Roy, A. Paulrath, and T. Kailath. Esprit - a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust. Speech Signal Processing*, 34(5):1340–1342, October 1986.

[24] P. Stoica, H. Li, and J. Li. Amplitude estimation of sinusoidal signals: Survey, new results, and an application. *IEEE Transactions on Signal Processing*, 48(2):338–345, February 2000.

[25] P. Stoica and R. Moses. *Spectral Analysis of Signals*. Prentice Hall, 2005.

[26] http://www.cs.ubc.ca/∼pai/ICAD07/.

[27] K. van den Doel, P. G. Kry, and D. K. Pai. FoleyAutomatic: Physically-Based Sound Effects for Interactive Simulation and Animation. In *Proceedings of ACM SIGGRAPH 2001*, Computer Graphics Proceedings, Annual Conference Series, pages 537–544, Aug. 2001.

[28] K. van den Doel and D. K. Pai. The sounds of physical shapes. *Presence*, 7(4):382–395, 1998.

[29] K. van den Doel, D. K. Pai, T. Adam, L. Kortchmar, and K. Pichora-Fuller. Measurements of perceptional quality of contact sound models. In *Proceedings of the International Conference on Auditory Display, Kyoto*, July 2002.