ATOMIC DECOMPOSITIONS OF AUDIO SIGNALS

Michael Goodwin [1] *Martin Vetterli* [2]

Department of Electrical Engineering and Computer Science
[1]Center for New Music and Audio Technologies (CNMAT)
University of California at Berkeley
[2]École Polytechnique Fédérale de Lausanne

ABSTRACT

Signal modeling techniques ranging from basis expansions to parametric approaches have been applied to audio signal processing. Motivated by the fundamental limitations of basis expansions for representing arbitrary signal features and providing means for signal modifications, we consider decompositions in terms of functions that are both signal-adaptive and parametric in nature. Granular synthesis and sinusoidal modeling can be viewed in this light; we interpret these approaches as signal-adaptive expansions in terms of time-frequency atoms that are highly correlated to the fundamental signal structures. This leads naturally to a discussion of the matching pursuit algorithm for deriving decompositions using overcomplete dictionaries of time-frequency atoms; specifically, we compare expansions using Gabor atoms and damped sinusoids. Such decompositions identify important signal features and provide parametric representations that are useful for signal coding and analysis–modification–synthesis.

1. SIGNAL DECOMPOSITIONS

Decompositions of signals in terms of elementary building blocks have been used in countless signal processing applications. In such decompositions, a signal $x[n]$ is represented as a linear combination of the form:

$$x[n] = \sum_{k=1}^{K} \alpha_k g_k[n] \qquad (1)$$

In a general analysis-synthesis framework, the expansion coefficients α_k are derived by the analysis and the expansion functions $g_k[n]$ are dictated by the synthesis, both in accordance with an underlying signal model. The set of coefficients and functions provide a representation of the signal; if the representation is compact or sparse, the decomposition indicates basic signal features and is generally useful for signal analysis and coding. It should be noted that compact representations tend to involve expansion functions that are highly correlated with the signal.

When the function set $g_k[n]$ constitutes a basis, a given signal has a unique expansion. This is of special interest for orthogonal cases such as wavelet or Fourier bases since the expansion coefficients can be independently derived and since fast computation algorithms such as the FFT are readily available. Such basis expansions, however, have a serious drawback in that a given basis is not well-suited for decomposing a wide variety of signals. As an example, consider the Fourier case: for a frequency-localized signal, a Fourier expansion is appropriately sparse and indicates the important signal features; for a time-localized signal, on the other hand, the Fourier representation does not readily provide information about the basic signal structure. This shortcoming results from the attempt to represent arbitrary signals in terms of a very limited set of functions. Better representations can be derived by using expansion functions that are signal-adaptive; this can be achieved by using adaptive wavelet packets or best basis methods, by parametric approaches such as the sinusoidal model, or by choosing the expansion functions from an overcomplete dictionary of time-frequency atoms. The latter two methods are of special interest here since they provide very flexible parametric representations.

1.1. Granular Synthesis

Granular synthesis is a technique in computer music which involves accumulating a large number of basic sonic components or *grains* to create a substantial acoustic event [1]. This approach is based on a theory of sound and perception that was first proposed by Gabor [2]; he suggested that any sound could be described using a quantum representation where each acoustic quantum or grain corresponds to a local time-frequency component of the sound. Furthermore, such descriptions are psychoacoustically appropriate given the time-frequency resolution tradeoffs and limitations observed in the auditory system.

In early efforts in granular music synthesis, artificial sounds were composed by combining thousands of parameterized grains [1]. Individual grains were generated according to synthetic parameters describing both time-domain and frequency-domain characteristics, for example time location, duration, envelope shape, and modulation. This method was restricted to the synthesis of artificial sounds, however, because the representation paradigm did not have an accompanying analysis capable of deriving granular decompositions of existing sounds.

Simple analysis techniques for deriving grains from real sounds were proposed in [3, 4]; the objective of such *granulation* approaches is to derive a representation of natural sounds that enables modifications such as time-scaling or pitch-shifting prior to resynthesis. The basic idea in these methods is to extract grains by applying time-domain windows to the signal. Each windowed portion of the signal is treated as a grain, and parameterized by its window function and time location. These grains can be realigned in time or resampled in various ways to achieve desirable signal modifications [3, 4]. Similar ideas have been developed in the speech processing community [5].

Grains derived by the time-windowing process can be interpreted as signal-dependent expansion functions that are highly correlated with the signal. If the grains are chosen judiciously, *e.g.* to correspond to pitch periods of a voiced sound, then the representation captures important signal structures and can as a result be useful for both coding and modification. Because of the complicated time structure of natural sounds, however, grains derived in this manner are generally difficult to represent efficiently and are thus not particularly applicable to signal coding. Nevertheless, this method is of interest because of its modification capabilities and its underlying signal adaptivity.

The time-windowed signal components derived by granulation are disparate from the fundamental acoustic quanta suggested by Gabor; time-windowing of the signal, while effective for modifications, is not an appropriate analysis for Gabor's time-frequency representation. This motivates both the following interpretation of the sinusoidal model as a granular analysis-synthesis and to a greater extent the subsequent consideration of signal decompositions based on parameterized time-frequency dictionaries.

1.2. Sinusoidal Modeling

In sinusoidal modeling, the signal is modeled as a sum of evolving sinusoids called *partials*. Analysis methods for this sum-of-partials model are generally frame-by-frame approaches based on the short-time Fourier transform (STFT); the parameters of the partials in a given frame are found by estimating the amplitude, frequency, and phase of the peaks in the Fourier spectrum of that frame [6, 7]. These parameters are then frame-rate representations of the time-varying amplitude and frequency *tracks* of the partials. The frame size and analysis stride of the STFT are chosen to give a reasonable tradeoff between the efficiency of the parameterization and its accuracy in modeling the signal.

Synthesis for the sinusoidal model can be achieved by summing the time-domain outputs of a bank of sinusoidal oscillators corresponding to the partials of the signal model. The output of a particular oscillator is dictated by amplitude and frequency control functions that are calculated in the synthesis process according to the parameters derived by the analysis. This involves two difficulties, line tracking and parameter interpolation, both of which arise because of frame-to-frame parameter differences for nonstationary signals. First, the synthesis must relate the parameter sets in adjacent frames to form partials that endure in time. This line tracking is generally done by coupling partials in adjacent frames if they are close in frequency; partials without appropriate pairings are accounted for as births or deaths, *i.e.* partials that are newly entering or leaving the signal. Second, after partial continuity is established, the frame-rate partial parameters must be interpolated to determine the sample-rate oscillator control functions. Typically, this interpolation is based on low-order polynomial models such as linear amplitude and quadratic frequency (cubic phase); the interpolation functions are constrained to meet amplitude, frequency, and phase matching conditions at the synthesis frame boundaries, which correspond in time to the centers of the analysis frames [6, 7].

The sinusoidal synthesis can be viewed as a sum of non-overlapping synthesis frames, each of which is a sum of partials. The reconstruction of the signal is given by

$$\hat{x}[n] \;=\; \sum_i \hat{x}_i[n] \;=\; \sum_i \sum_{q=1}^{Q_i} A_{q,i}[n] \cos \Theta_{q,i}[n] \qquad (2)$$

where $A_{q,i}[n]$ and $\Theta_{q,i}[n]$ are functions derived by interpolating the analysis parameters as described above; i is a frame index and q is a partial index. Note that each of the modulated sinusoids in this expression is time-localized to a synthesis frame and frequency-localized according to its quadratic frequency function. In this light, the sinusoidal model can be interpreted as a decomposition in terms of linear-amplitude, cubic-phase sinusoidal atoms. These atoms are generated directly from the analysis data by the synthesis interpolation; the expansion functions in the sinusoidal model are thus signal-dependent, and the time-frequency representation is signal-adaptive.

Sinusoidal modeling derives an accurate reconstruction for signals that vary slowly with respect to the frame rate. Variations that occur on shorter time scales are not well-modeled. This difficulty is clearly explained by the atomic interpretation of the sinusoidal model; any rapid transient is spread out across a synthesis frame because of the fixed time resolution of the sinusoidal model expansion functions. In short, these functions are inadequate for rapidly varying signals. This situation can be remedied by applying a multiresolution framework to the sinusoidal model, either by subband filtering followed by sinusoidal modeling of the channel signals with long frames for low-frequency bands and short frames for high-frequency bands, or by using a time-varying segmentation with short frames near transients and long frames for stationary behavior [8]. Such methods admit expansion functions with an appropriate variety of time supports into the decomposition; this results in a more accurate signal representation.

Sinusoidal modeling has found many applications in speech and audio coding and analysis-synthesis. This is primarily because the representation in terms of sinusoidal parameters is efficient and readily allows for such desirable transformations as time-scaling, pitch-shifting, and cross-synthesis [6, 7]. The efficiency results from the signal-adaptivity of the expansion functions, and the modification capabilities arise because the representation is parametric in nature. In effect, if the sinusoidal parameters are modified, a new set of expansion functions are derived based on those modifications. This robustness to modification results directly from the parametric nature of the representation. Basis-type expansions in terms of fixed vectors do not exhibit a similar flexibility; *e.g.* if the coefficients in a wavelet filterbank expansion are modified, aliasing is introduced. Furthermore, it is not clear how to achieve a desired perceptually-oriented modification such as pitch-shifting for an arbitrary expansion.

The preceding discussions argued that granular synthesis and sinusoidal modeling can both be interpreted as time-frequency atomic methods, and that these approaches are useful due to the signal-adaptivity of the representation and their inherent parametric structures. Another way to achieve a parametric signal-adaptive representation is to choose the expansion functions from an overcomplete dictionary of parameterized time-frequency atoms. Expansions based on arbitrary dictionaries can be derived using the matching pursuit algorithm, which is described in the next section. Then, parametric dictionaries that readily allow for signal modifications will be discussed.

2. MATCHING PURSUIT

Matching pursuit is a recently proposed algorithm for deriving signal decompositions in terms of expansion functions chosen from a *dictionary* [9]. To achieve a signal-adaptive representation, an *overcomplete* dictionary is used, meaning that the dictionary contains a basis for the signal space plus additional functions. This overcompleteness or redundancy implies that the dictionary elements, or *atoms*, exhibit a wide range of behaviors, and can thus provide better decompositions of a wide range of signals than a basis expansion. This approach is adaptive in the sense that the algorithm chooses the appropriate atoms from the dictionary to decompose a particular signal.

Matching pursuit refers specifically to a greedy iterative algorithm for determining an expansion given a signal and a dictionary of atoms. At each stage of the iteration, the atom that best approximates the signal is chosen; then the weighted contribution of this atom to the signal is subtracted and the iteration proceeds on the residual. Using the two-norm as the approximation metric, the task at the i-th stage of the algorithm is to find the atom $g_{m(i)}[n]$ that minimizes the two-norm of the residual signal

$$r_{i+1}[n] \;=\; r_i[n] - \alpha_i g_{m(i)}[n] \qquad (3)$$

where α_i is a weight that describes the contribution of the atom to the signal, *i.e.* the expansion coefficient, and $m(i)$ is

the dictionary index of the atom; the iteration begins with $r_1[n] = x[n]$, the original signal. The solution for α_i and $g_{m(i)}[n]$ follows from the orthogonality principle; treating the signals as column vectors, the two-norm of the residual r_{i+1} is a minimum if it is orthogonal to the atom:

$$\langle r_i - \alpha_i g_{m(i)}, g_{m(i)} \rangle = (r_i - \alpha_i g_{m(i)})^H g_{m(i)} = 0 \quad (4)$$

$$\Longrightarrow \alpha_i = \frac{\langle g_{m(i)}, r_i \rangle}{\langle g_{m(i)}, g_{m(i)} \rangle} = \langle g_{m(i)}, r_i \rangle$$

where the last step follows from restricting the atoms to be unit-norm. Then, the two-norm $\langle r_{i+1}, r_{i+1} \rangle$ of the error is

$$\langle r_i, r_i \rangle - \frac{|\langle g_{m(i)}, r_i \rangle|^2}{\langle g_{m(i)}, g_{m(i)} \rangle} = \langle r_i, r_i \rangle - |\alpha_i|^2. \quad (5)$$

This energy is minimized by choosing the atom $g_{m(i)}$ that has the largest magnitude correlation with the signal r_i, and the expansion coefficient for that atom is $\langle g_{m(i)}, r_i \rangle$.

In deriving a signal decomposition, the matching pursuit iteration is continued until the residual energy is below some threshold, or until some other halting criterion is met. After I iterations, the algorithm gives the signal estimate

$$x[n] \approx \hat{x}[n] = \sum_{i=1}^{I} \alpha_i g_{m(i)}[n], \quad (6)$$

Note that the expansion functions here are highly correlated with the signal as in the methods of section 1. The mean-squared error of this approximate decomposition, namely the energy of the residual $x[n] - \hat{x}[n]$, converges to zero as the number of iterations approaches infinity [9]. The convergence property of this successive approximation implies that I iterations will provide a reasonable I-term decomposition of the signal; global optimality, however, is not insured because of the nature of the algorithm. Determining the globally optimal I-term expansion based on an overcomplete dictionary requires finding the minimum error over all I-dimensional dictionary subspaces, which is not computationally feasible for large I [10].

To enable representation of a wide range of signal features, large dictionaries are used in the matching pursuit algorithm. The computation of the correlations $\langle g, r_i \rangle$ is thus intensive. As noted in [9], however, the computation can be drastically reduced using an update formula derived from equation 3; the correlations at stage $i+1$ are directly related to the correlations at stage i by the equation

$$\langle g, r_{i+1} \rangle = \langle g, r_i \rangle - \alpha_i \langle g, g_{m(i)} \rangle \quad (7)$$

where the only new computation required for the correlation update is the dictionary cross-correlation term $\langle g, g_{m(i)} \rangle$, which can be precomputed and stored.

3. TIME-FREQUENCY DICTIONARIES

Matching pursuit can be viewed as a method of finding sparse approximate solutions to inverse problems [11]. This is equivalent to deriving sparse signal decompositions in terms of arbitrary dictionaries of expansion functions. Such decompositions are not particularly useful, however, unless the functions correspond to relevant signal structures. Thus, overcomplete dictionaries consisting of atoms that exhibit a wide range of localized time and frequency behaviors are of significant interest; decomposition in terms of such atoms provides an adaptive time-frequency representation of a signal [9]. Such localized time-frequency atoms correspond to the perceptually motivated quanta introduced by Gabor. Matching pursuit using such atoms thus provides a compatible analysis method for granular synthesis, especially since the atoms in typical time-frequency dictionaries are parameterized in such a way that signal modifications can be achieved.

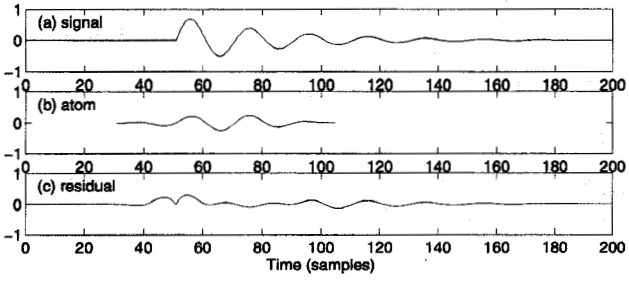

Figure 1. *(a) a damped sinusoidal signal, (b) the optimal first atom chosen from a symmetric Gabor dictionary, and (c) the residual; note the artifact near the onset time.*

3.1. Gabor Atoms

The literature on matching pursuit has focused largely on applications using dictionaries of Gabor atoms; these are appropriate expansion functions for time-frequency signal decompositions [9]. Such atoms are scaled, modulated, and translated versions of a single window function:

$$g_{\{s,\omega,\tau\}}[n] = \frac{1}{\sqrt{s}} g\left(\frac{n-\tau}{s}\right) e^{j\omega n} \quad (8)$$

In this notation, $g(t)$ is a unit-norm window function from which the atoms are derived; each atom is indexed in the dictionary by a parameter set $\{s, \omega, \tau\}$. This parametric structure allows for a simple description of a specific dictionary and provides modification capabilities; this is a two-fold advantage with respect to dictionaries of arbitrary vectors. A Gabor dictionary, which includes Fourier and wavelet-like bases, is highly overcomplete when the scale, modulation, and translation parameters are not tightly restricted; such overcompleteness, especially when coupled with a nonlinear analysis like matching pursuit, yields signal-adaptive representations [12].

In applications of Gabor functions, $g(t)$ is typically an even-symmetric window. The associated dictionaries thus consist of atoms that exhibit symmetric time-domain behavior. This is problematic for representing asymmetric signal features such as transients, which occur frequently in natural signals such as music. Figure 1(a) shows a typical transient from linear system theory, the damped sinusoid; the first stage of a matching pursuit based on symmetric Gabor functions chooses the atom shown in Figure 1(b). This atom matches the frequency behavior of the signal, but its time-domain symmetry results in a *pre-echo* artifact in the residual as shown in Figure 1(c). The residual has energy before the onset of the original signal, which the matching pursuit algorithm must then remove at subsequent stages. One approach to this problem is the high-resolution matching pursuit algorithm suggested in [13, 14], where symmetric atoms are still used but the correlation metric is modified so that atoms that introduce such artifacts are not chosen for the decomposition. Another approach is to use a dictionary of asymmetric atoms such as damped sinusoids.

3.2. Damped Sinusoids

The common occurrence of damped oscillations in natural signals is sufficient justification for considering damped sinusoids as building blocks in signal decompositions. This matching pursuit application is further motivated in that damped sinusoids are better suited than symmetric Gabor atoms for representing transients. Like the atoms in a general Gabor dictionary, damped sinusoidal atoms can be indexed by characteristic parameters; the damping factor a, modulation frequency ω, and start time τ specify these atoms:

$$g_{\{a,\omega,\tau\}}[n] = S\, a^{(n-\tau)} e^{j\omega n} u[n-\tau] \quad (9)$$

where S is a scaling factor needed to satisfy the unit-norm requirement. It should be noted that damped sinusoidal

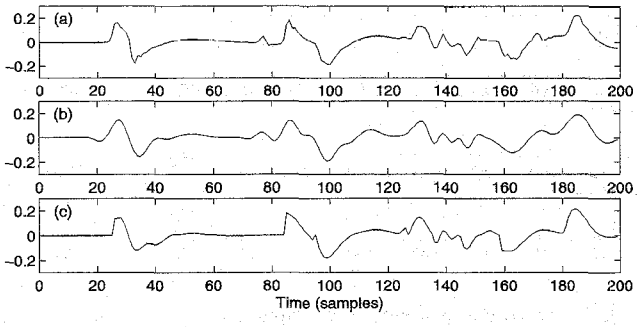

Figure 2. *(a) A gong attack, and decompositions using (b) 10 symmetric Gabor atoms, and (c) 10 damped sinusoids.*

atoms can be interpreted as Gabor functions derived from a one-sided exponential window; they are just differentiated from typical Gabor atoms by their asymmetry. Also, the atomic structure is more readily indicated by a damping factor rather than a scale parameter, so the dictionary index set is different than in the general Gabor case.

For a dictionary of damped sinusoids, the structure of the signal–atom correlations allows for reduction in the computation irrespective of equation 7 as well as an intuitive interpretation of the pursuit algorithm. The correlation computation over the time index is simplified based on the exponential structure of the atoms, which results in a recursion relation between correlations at neighboring times:

$$\rho(a,\omega,\tau-1) \;=\; ae^{-j\omega}\rho(a,\omega,\tau) \;+\; \ldots \qquad (10)$$

where the omitted terms are corrections to account for truncation of the atoms [11]. This relation is simply a one-pole filter; it suggests interpreting the exhaustive correlation computation as an application of the signal to a dense grid of one-pole filters, which are the *matched filters* for the dictionary atoms. A further simplification can be achieved if the dictionary has a harmonic structure; for any dictionary of harmonically modulated atoms, the FFT can be used to compute correlations over the frequency index [11].

The dictionaries discussed consist of complex atoms. This is not problematic for applications with real signals since the pursuit algorithm can be modified to derive real expansions of real signals based on a complex dictionary [9, 11].

Figure 2 shows a comparison of decompositions using symmetric atoms and damped sinusoids. The dictionaries are designed for a fair comparison; the errors in the two pursuits exhibit similar convergence behavior as shown in Figure 3(a). The symmetric decomposition, however, introduces a pre-echo before the signal onset as depicted in Figure 2(b). Later iterations of the pursuit are devoted to removing this artifact; this is characterized in Figure 3(b), which shows the mean-squared energy of the reconstruction pre-echo as a function of the number of iterations. As demonstrated in Figure 2(c), the decomposition with damped sinusoids does not introduce a pre-echo.

For the signal in Figure 2(a), the dictionary of damped sinusoids outperforms the symmetric dictionary with respect to representing the signal onset, which is particularly important in music perception [6]. In general, however, the symmetric decomposition does provide a reasonable representation of the signal. This points to the conclusion that the choice and design of dictionaries inherently depends on very specific properties desired for the representation. For instance, the symmetric decomposition yields a very smooth reconstruction in the early stages of the pursuit, which is appropriate in some signal approximation scenarios. In arbitrary cases, hybrid dictionaries may well prove most useful for representing basic signal structures. This issue of dictionary design is an open and extremely relevant question.

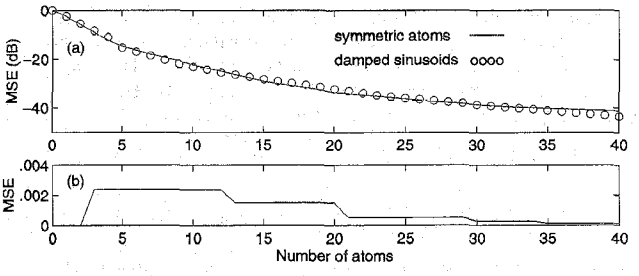

Figure 3. *(a) The mean-squared reconstruction error using symmetric atoms and damped sinusoids, and (b) the energy in the pre-echo of the symmetric-atom decomposition.*

4. FUTURE WORK

The use of matching pursuit for signal analysis has been explored in the literature [9, 13, 14]. In future work we plan to investigate applications to high-quality audio coding and to formalize the modification capabilities mentioned here. In light of its relationship to the other useful methods discussed in this paper, we anticipate that the matching pursuit approach for deriving parametric signal-adaptive time-frequency atomic decompositions will be effective for audio applications. For instance, we expect these atomic models to facilitate the use of psychoacoustic masking principles to improve compression performance.

REFERENCES

[1] C. Roads. Introduction to granular synthesis. *Computer Music Journal*, 12(2):11–13, Summer 1988.

[2] D. Gabor. Acoustical quanta and the theory of hearing. *Nature*, 159(4044):591–594, May 1947.

[3] D. Jones and T. Parks. Generation and combination of grains for music synthesis. *Computer Music Journal*, 12(2):27–34, Summer 1988.

[4] B. Truax. Discovering inner complexity: Time shifting and transposition with a real-time granulation technique. *Computer Music Journal*, 18(2):38–48, 1994.

[5] E. Moulines and J. Laroche. Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech Communication*, 16:175–205, 1995.

[6] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, Winter 1990.

[7] R. McAulay and T. Quatieri. Speech analysis / synthesis based on a sinusoidal representation. *IEEE-ASSP*, 34(4), August 1986.

[8] P. Prandoni, M. Goodwin, and M. Vetterli. Optimal segmentation for signal modeling and compression. *ICASSP Proceedings*, 1997.

[9] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE-SP*, 41(12):3397–3415, December 1993.

[10] G. Davis. *Adaptive Nonlinear Approximations*. PhD thesis, New York University, 1994.

[11] M. Goodwin. Matching pursuit with damped sinusoids. *ICASSP Proceedings*, 1997.

[12] V. Goyal, N. Thao, and M. Vetterli. Quantized overcomplete expansions in $\Re^N$: Analysis, synthesis and algorithms. *IEEE-IT*. To appear.

[13] S. Jaggi et al. High resolution pursuit for feature extraction. Tech. Rep. LIDS-P-2371, MIT, Nov. 1996.

[14] R. Gribonval et al. Analysis of sound signals with high resolution matching pursuit. *Proceedings of TFTS*, pp. 125–128, June 1996.