euronoise

A comparison of molecular approaches for generating sparse and structured multiresolution representations of audio and music signals

Bob Sturm[a], John Shynk[a], Aaron McLeran[b], Curtis Roads[b] and Laurent Daudet[c]

[a]University of California, Box 117, Department of Electrical and Computer Engineering, Santa Barbara, CA 93106, USA
[b]University of California, Media Arts and Technology Program, Santa Barbara, CA 93106, USA
[c]UPMC Univ Paris 06, LAM / IJLRA, 11 rue de Lourmel, 75015 Paris, France
boblsturm@ece.ucsb.edu

We compare the characteristics and performance of joint (single-step) and sequential (two-step) approaches for creating sparse and structured acoustic signal representations derived using overcomplete methods (OMs). A joint approach, such as molecular matching pursuit (MMP), attempts to find coherent structures in a signal as part of the decomposition process, while a sequential approach, such as agglomerative clustering (AC), attempts to find coherent structures after the signal decomposition. We review each approach, and examine their performance using real audio and music signals.

1 Introduction

Acoustic signals exhibit a wide variety of structures, such as the impulses and resonances of musical instruments. Such instruments possess unique and identifiable structures consisting of general high-level features, such as pitch, vibrato, and timbre, and specific low-level characteristics, such as attack time and harmonic relationships. These structures can be viewed as "content" that one might use for applications of analysis, discrimination, and transformation. To work at such diverse levels of detail, one needs a method that is capable of efficiently and meaningfully representing content in a flexible manner. Sparse approximations, or overcomplete methods (OMs), aim to provide such representations.

OMs attempt to overcome the limitations of orthogonal signal transformations (e.g., Fourier) by generalizing the transformation process to a decomposition based on an arbitrarily specified set (*dictionary*) of time-localized functions (*atoms*). Instead of assuming a signal will be well-described by a specific basis, OMs use dictionaries of atoms that can be specified without orthogonality restrictions. An atom can take any shape and scale, and might even be tuned to different structures expected in a signal, such as transients and tonals [1], or to particular musical instruments [2]. Compared to the weights of a localized Fourier basis, an atomic representation resulting from OMs can manifest a more informative representation of the structures in a signal.

While OMs can significantly reduce the dimensionality of a signal, the significance of each atom to particular content of that signal may not be clear. For instance, the relationship of an atom of a particular scale to content at a larger scale may not be evident. We have thus sought ways to make more clear these relationships by creating sparse and structured representations through OMs using *molecules* of atoms [1–3]. The basic idea of these methods is to group atoms into structurally significant components. In a *joint* approach, the decomposition process adapts the atom selection criteria to local aspects of the signal, such as tonalicity. In a *sequential* approach, atoms of a completed decomposition are grouped together using rules and measures of similarity, such as time-frequency (TF) overlap. After reviewing OMs, we describe these two approaches, and then discuss their application to real audio signals.

2 Overcomplete Methods

Consider a complex K-dimensional vector space $\mathcal{X}_K \in \mathbb{C}^K$ with an inner product between two vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}_K$ defined as $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. The ℓ^2-norm of any $\mathbf{x}_i \in \mathcal{X}_K$ is given by $||\mathbf{x}_i||_2 = \sqrt{\mathbf{x}_i^H \mathbf{x}_i}$ where ^{H} denotes complex conjugate transpose. Let the dictionary be described by the set $\{\mathbf{d}_i \in \mathcal{X}_K : ||\mathbf{d}_i||_2 = 1\}$, which can be expressed in matrix form as $\mathbf{D} = [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]_{K\times N}$. Now, given $\mathbf{x} \in \mathcal{X}_K$, a solution to the following problem is desired:

$$\min f\Big(C(\mathbf{s}),\, D(\mathbf{x},\mathbf{r})\Big) \text{ subject to } \mathbf{x} = \mathbf{D}\mathbf{s} + \mathbf{r} \qquad (1)$$

where $f(\cdot)$ is composed of a cost function $C(\mathbf{s})$ of the dictionary weights $\mathbf{s} \in \mathbb{C}^N$, and a distortion measure $D(\mathbf{x},\mathbf{r})$ using the original signal and an error $\mathbf{r}$. This joint-minimization is often contradictory in that decreasing one quantity can increase the other. One can thus fix either function and minimize the other. Matching pursuit (MP) [4] is an iterative descent OM that finds good solutions to (1) quickly by minimizing the residual energy at each step. Its solutions, however, are often suboptimal with respect to sparsity [5], which may or may not be important, depending on the intended application of an approximation.

MP proceeds as follows. Given a signal $\mathbf{x} \in \mathbb{C}^K$ and dictionary $\mathbf{D} \in \mathbb{C}^{K\times N}$, the output at the nth iteration is the representation $\{\mathbf{H}(n), \mathbf{a}(n), \mathbf{r}(n)\}$ such that $\mathbf{x} = \mathbf{H}(n)\mathbf{a}(n) + \mathbf{r}(n)$ where the columns of $\mathbf{H}(n) = [\mathbf{h}_0|\mathbf{h}_1|\cdots|\mathbf{h}_{n-1}]_{K\times n}$ are atoms selected from $\mathbf{D}$, and $\mathbf{a}(n) = [a_0, \ldots, a_{n-1}]^T$ contains their weights. The nth order residual is $\mathbf{r}(n) = \mathbf{x} - \mathbf{H}(n)\mathbf{a}(n)$ where n refers to the decomposition iteration, and is not the same as the time index k of the signal. The nth iteration of MP finds a new atom and its weight by

$$\mathbf{h}_n = \arg\max_{\mathbf{d}\in\mathbf{D}} |\langle \mathbf{d}, \mathbf{r}(n)\rangle| \qquad (2)$$

$$a_n = \langle \mathbf{h}_n, \mathbf{r}(n)\rangle, \qquad (3)$$

with $\mathbf{r}(0) \equiv \mathbf{x}$. After updating $\mathbf{H}(n+1) = [\mathbf{H}(n)|\mathbf{h}_n]$ and $\mathbf{a}(n+1) = [\mathbf{a}^T(n), a_n]^T$, the new residual is given by $\mathbf{r}(n+1) = \mathbf{x} - \mathbf{H}(n+1)\mathbf{a}(n+1) = \mathbf{r}(n) - a_n\mathbf{h}_n$, and the process repeats until some stopping criteria are met. When $n = 1$, MP is equivalent to the codeword selection of gain-shape vector quantization [4, 6]. OMs can be viewed as generalizations of this process to finding the best n vectors from $\mathbf{D}$ with respect to a cost/distortion criterion [7].

3 Molecular Representations

Figure 1 shows an audio signal represented in the TF plane. The spectrogram was created with a 46 ms Hann window and a uniform hop of 2 ms. The sparse approximation was found using MP and a multiscale dictionary of modulated and translated Hann windows (decomposed until the signal-to-residual energy ratio (SRR) = 30 dB). While the spectrogram contains 206,720 values, the sparse approximation consists of only 2,456 terms. To extract structural information from these TF representations, one must relate each element to its neighbors and to the content they represent together.

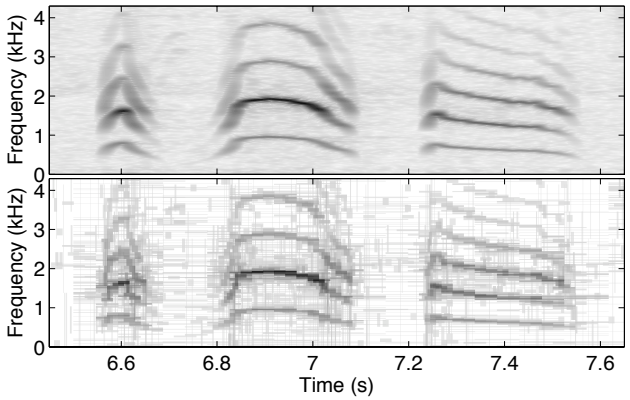

Figure 1: Segment of a bird call in time-frequency (TF) domain via a spectrogram (top), and a superposition of TF tiles of atoms found by MP (bottom).

Assuming a model of sines plus noise, the McAulay-Quatieri algorithm (MQA) [8] performs such an analysis using the spectrogram, distilling it into a set of parameters controlling sinusoids and noise sources to synthesize the original signal. It can clearly be seen, however, that working in the low-dimensional space of a sparse approximation can provide a considerable advantage because the atoms embody a high level of significance with respect to signal structures. For instance, in Fig. 1, longer atoms represent relatively stationary content.

Using OMs, we want to find and delimit significant structures in acoustic signals by grouping atoms into molecules that have particular functions in the signal. This can provide sparse representations having many different levels of content, from a high level of coarse phrasing or source discrimination, to a middle level of individual notes or voices, to a low level of transients and partials. We now review two methods for building such sparse and structured representations: a joint (one-step) approach where molecules are constructed during the decomposition process, and a sequential (two-step) approach where molecules are built after the decomposition process.

3.1 Joint Approach: Molecular MP

Molecular MP (MMP) [1] is an OM that decomposes a signal by extracting groups of atoms that serve either a tonal or a transient function in the signal. It accelerates the MP decomposition process by taking advantage of the mutually exclusive properties of each type of structure. Each atom of a molecule is found in relation to others according to specified rules for each structure type. The dictionary used in [1] is a union of two sets of atoms: windowed cosines (C) and dyadic wavelets (W).

The tonal contents of a signal are represented solely by atoms selected from C, which is built from a modified discrete cosine transform (MDCT) basis [9] with constant scale $s > 0$ and hop size $s/2$, a unique translation index $p \in \mathbb{Z}$, and modulation frequency index $l = 0, 1, \ldots, s/2$. Each MDCT atom is given by

$$h_{\mathsf{C}}(k; s, l, p) = Y_{s,l,p} f(k - ps/2; s) \times \cos\left[\frac{2\pi}{s}\left(k + \frac{s/2+1}{2} - ps/2\right)\left(l + \frac{1}{2}\right)\right] \quad (4)$$

where the window $f(k; s)$ satisfies a perfect time-domain aliasing cancelation constraint, such as the sine window

$$f(k; s) = \begin{cases} \sin\left(\frac{\pi}{s}(k + 0.5)\right), & k = 0, 1, \ldots, s-1 \\ 0, & \text{else.} \end{cases} \quad (5)$$

Finally, $Y_{s,l,p}$ is a scalar that makes the atom have unit norm. Ten atoms of C are shown in Fig. 2 for $s = 1024$.

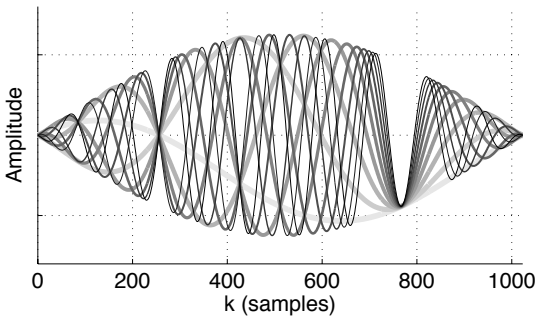

Figure 2: MDCT atoms $h_{\mathsf{C}}(k; 1024, l, 0)$, $l \in \{0, \ldots, 9\}$.

The transient contents of a signal are represented solely by atoms selected from W, which are defined by dilating and translating a generating wavelet $w_0(t)$

$$g_{\mathsf{W}}(t; j, u) = \frac{1}{\sqrt{2^j}} w_0\left(\frac{t-u}{2^j}\right) \quad (6)$$

with the property that each wavelet with scale index $j > 1$ can be described as a linear combination of two "children" wavelets of scale index $j - 1$. The set W thus forms a family of translated dyadic wavelet trees with maximum scale J. Figure 3 shows the wavelets generated using a Daubechies filter of length 4 [10] for scale indices $j \in \{1, \ldots, 9\}$.

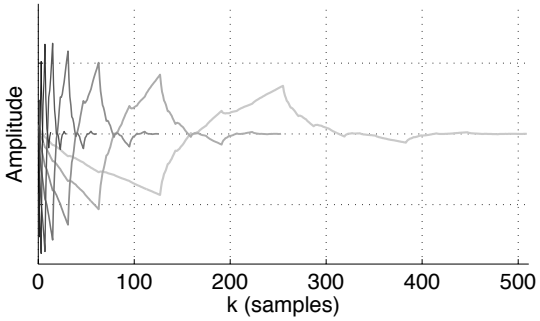

Figure 3: Wavelet atoms $h_{\mathsf{W}}(k; j, 0)$, $j \in \{1, \ldots, 9\}$.

3.1.1 Measuring Tonal and Transient Content

Let $\mathbf{h}_{s,l,p} \leftarrow h_{\mathsf{C}}(k; s, l, p)$ and $\mathbf{g}_{j,u} \leftarrow g_{\mathsf{W}}(k; j, u)$ be vector forms of the atoms in each subdictionary, and define the following functions embodying the inner products of the nth-order residual $\mathbf{r}(n)$:

$$\beta(s, l, p) \triangleq \langle \mathbf{h}_{s,l,p}, \mathbf{r}(n) \rangle, \; \alpha(j, u) \triangleq \langle \mathbf{g}_{j,u}, \mathbf{r}(n) \rangle. \quad (7)$$

At iteration n, the values in (7) are evaluated to determine whether a tonal or transient molecule is extracted. Due to the construction of $\mathsf{C} \cup \mathsf{W}$, these quantities can be calculated quickly using the MDCT, and a multirate perfect reconstruction filterbank [10].

The strength of tonal content is gauged using the *local tonality index*, defined as

$$\mathcal{T}(s, l, p) \triangleq \frac{1}{W} \sum_{i=0}^{W-1} \mathcal{S}_{s,l,p+i}\{\mathbf{r}(n)\} \quad (8)$$

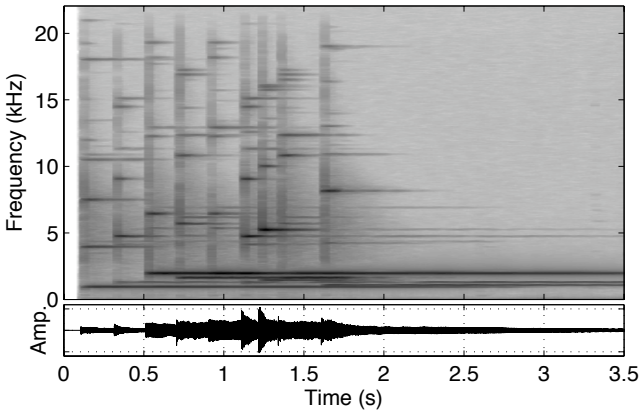

Figure 4: Logarithm of local tonality index for a glockenspiel calculated using $s = 1024$ and $W = 5$.

where $W > 0$, and the localized "pseudo-spectrum" of $\mathbf{r}(n)$ is defined as

$$\mathcal{S}^2_{s,l,p}\{\mathbf{r}(n)\} \triangleq \beta^2(s,l,p) + [\beta(s,l+1,p) - \beta(s,l-1,p)]^2 \quad (9)$$

for $l = 0, 1, \ldots, s/2-1$, and $\beta(s,-1,p) = \beta(s,s/2,p) = 0$. The pseudo-spectrum gauges the spread of energy into adjacent frequency bins $l+1$ and $l-1$. Thus, (8) is a W-order causal moving average of (9). $\mathcal{T}(s,l,p)$ quantifies the strength of a tonal component centered on frequency l over a duration of W frames, independent of its phase. Figure 4 shows the local tonality index using $s = 1024$ and $W = 5$ for a monophonic musical signal.

To determine the strength of transient content, MMP uses the *modulus of regularity*, defined as

$$\kappa(2u) \triangleq \frac{1}{J} \sum_{(j,u)\in I_{2u}} |\alpha(j,u)| \quad (10)$$

where I_{2u} contains the scales and translations of wavelets related to the smallest scale wavelet translated to $2u$. The measure $\kappa(2u)$ is thus the average magnitude of the inner products of $\mathbf{r}(n)$ with all the wavelets in a connected tree branch from the smallest scale ($j = 1$) wavelet, to the largest scale J translated to $\lfloor 2u/2^J \rfloor$.

3.1.2 MMP Decomposition Process

Each iteration considers $\mathcal{T}(s,l,p)$ and $\kappa(2u)$, computed from the residual signal, to extract either a tonal or transient molecule based on the following decision:

$$\max_{s,l,p} \mathcal{T}(s,l,p) \underset{\text{transient}}{\overset{\text{tonal}}{\gtrless}} \max_{u} \kappa(2u). \quad (11)$$

The decomposition process stops when both sides are less than some small $\epsilon > 0$, which signifies that the residual no longer contains tonal and transient content.

A tonal molecule can be arbitrarily long, and is built around the frequency index l_0 of the maximum value of $\mathcal{T}(s,l,p)$. MMP searches for atoms within one frequency index of l_0, i.e., $[l_0-1, l_0, l_0+1]$, and the translation indices $[p_{\text{start}}, p_{\text{end}}]$, found using endpoint detection on $\mathcal{S}_{s,l_0,p}\{\mathbf{r}(n)\}$. MMP subsumes into a set at step n the parameters and coefficients of the atoms in C that have a projection magnitude exceeding ϵ for $p \in \{p_{\text{start}}, \ldots, p_{\text{end}}\}$ and $l \in \{l_0-1, l_0, l_0+1\}$, i.e.,

$$\mathcal{M}^{\mathsf{C}}_n = \{s,l,p,\beta(s,l,p) : |\beta(s,l,p)| > \epsilon\}. \quad (12)$$

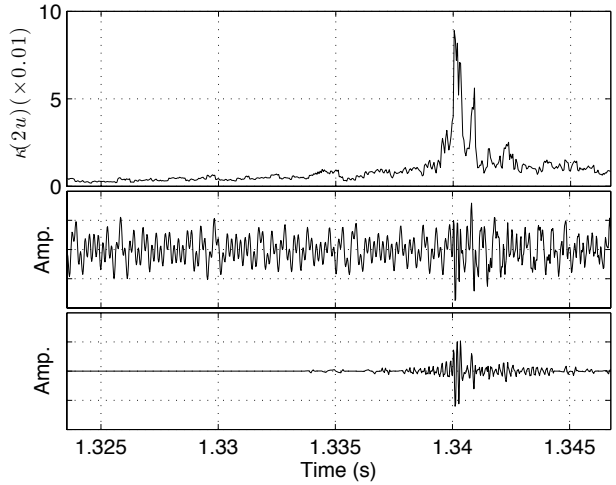

Figure 5: Modulus of regularity (top) for a glockenspiel, using $J = 9$ and $\epsilon = 0.02$. Relevant signal segment (middle). Extracted transient (bottom).

MMP builds a transient molecule by pruning the wavelet tree associated with the maximum modulus of regularity. Figure 5 shows an example for a wavelet tree containing transient content in the signal seen in Fig. 4. Each branch of this fully connected tree is pruned starting at the smallest scale ($j = 1$) until $|\alpha(j,u)| < \epsilon$ for some $j < J$. The remaining fully connected tree thus constitutes a transient molecule at step n:

$$\mathcal{M}^{\mathsf{W}}_n = \{j, u, \alpha(j,u) : |\alpha(j,u)| > \epsilon\}. \quad (13)$$

3.2 Sequential Approach: AC

Another approach to building a structured representation is through agglomerative clustering (AC) of a sparse decomposition [3]. This method builds molecules based on measures of similarity between atoms. In this way a sparse approximation is structured according to rules specific to particular content. The dictionary used in [3] is a set C of windowed cosines of multiple scales s

$$h_{\mathsf{C}}(k; s, \omega, u, \phi) = Y_{s,\omega} f(k-u; s) \cos(\omega[k-u] + \phi) \quad (14)$$

where $f(k;s)$ is a Hann window with support s, u is a translation, $0 \le \omega \le \pi/2$ and $0 \le \phi < 2\pi$ are modulation parameters, and $Y_{s,\omega}$ is a scalar making each atom have unit norm. The atom parameters are discretized in the following way: $s \in \{2^r, r = 1, 2, \ldots, 14\}$, $u = zs/2$ for $z \in \mathbb{Z}$, and $\omega \in \{2\pi l/s, l = 0, 1, \ldots, s/2\}$. Phase ϕ is not discretized since complex atoms are used.

3.2.1 Measuring Atomic Similarity

The basic principle of AC is that a pair of atoms belong to the same molecule if they are sufficiently similar. A simple measure of similarity between a pair of atoms $\{\mathbf{h}_{\gamma_i}, \mathbf{h}_{\gamma_j}\}$, where $\mathbf{h}_\gamma \leftarrow h_{\mathsf{C}}(k;\gamma)$ is the vector form of an atom in C, and $\gamma = \{s, \omega, u, \phi\}$, is the magnitude of their analytic inner product

$$\rho_{ij} \triangleq \left| \langle \tilde{\mathbf{h}}_{\gamma_i}, \tilde{\mathbf{h}}_{\gamma_j} \rangle \right| \quad (15)$$

where $\tilde{\mathbf{h}}_\gamma$ is the analytic form of $\mathbf{h}_\gamma$, e.g., for (14)

$$\tilde{h}_{\mathsf{C}}(k;\gamma) = Y_{s,\omega} f(k-u;s) e^{\sqrt{-1}(\omega[k-u]+\phi)}. \quad (16)$$

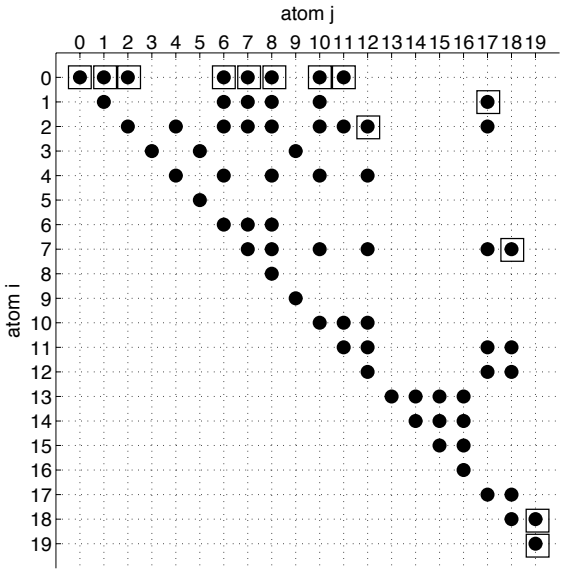

Figure 6: Adjacency matrix and agglomeration process. Boxes contain unique atoms included in the molecule.

Thus, a pair of atoms is similar if $\rho_{ij} \geq \rho_{\min}$, where the correlation threshold $0 \leq \rho_{\min} \leq 1$. In effect, this rule implies that atoms are similar if they sufficiently overlap in time and frequency.

While the previous agglomeration rule is reasonable for pairs of atoms that have a narrow bandwidth, e.g., large scale atoms, it may not be useful for short-scale atoms. Such atoms can be grouped by defining similarity using the difference between their center-times, i.e.,

$$\mu_{ij} \triangleq \mu_{\min} - \left|(u_i + s_i/2) - (u_j + s_j/2)\right| \qquad (17)$$

where the distance threshold $\mu_{\min} > 0$. If $\mu_{ij} \geq 0$, then the pair of atoms are similar.

3.2.2 Agglomerative Clustering Process

Through the similarity measures described above, AC finds and delimits tonal and transient content in an nth-order sparse approximation $\{\mathbf{H}(n), \mathbf{a}(n), \mathbf{r}(n)\}$. Assuming that large-scale atoms represent tonal content, and small-scale atoms represent transient content, the atoms contained in $\mathbf{H}(n)$ are separated into two sets based on their scales: $\mathcal{H}_{>\sigma} = \{h_{\mathsf{C}}(k;\gamma) : s > \sigma\}$ for atoms with scales $> \sigma$, and $\mathcal{H}_{\leq\sigma} = \{h_{\mathsf{C}}(k;\gamma) : s \leq \sigma\}$ otherwise. Since these two sets are disjoint, i.e., $|\mathcal{H}_{>\sigma}| + |\mathcal{H}_{\leq\sigma}| = n$, tonal and transient molecules can be built in parallel.

From the first set $\mathcal{H}_{>\sigma}$, AC constructs a binary *adjacency matrix* $\mathbf{A}(n)$ with entries assigned as follows:

$$a_{ij} = \begin{cases} 1, & \rho_{ij} \geq \rho_{\min},\ 1 < j \leq |\mathcal{H}_{>\sigma}|,\ i \leq j \\ 0, & \text{else.} \end{cases} \qquad (18)$$

This upper-triangular matrix specifies which pairs of atoms are sufficiently similar. By traversing the elements of $\mathbf{A}(n)$, AC agglomerates atom pairs into tonal molecules. This is done until all nonzero entries of the adjacency matrix have been searched. A new molecule is then started using the remaining atoms. This process is the same for constructing transient molecules, but using a binary adjacency matrix $\mathbf{B}(n)$ having entries

$$b_{ij} = \begin{cases} 1, & \mu_{ij} \geq 0,\ 1 < j \leq |\mathcal{H}_{\leq\sigma}|,\ i \leq j \\ 0, & \text{else.} \end{cases} \qquad (19)$$

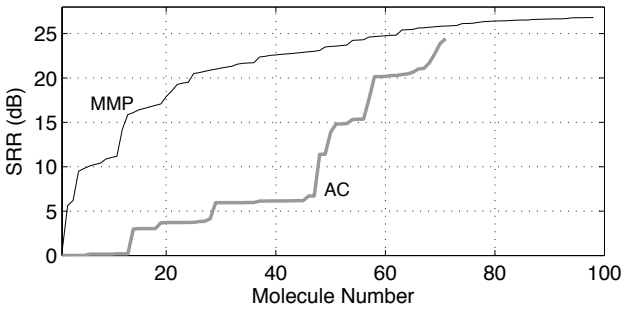

Figure 7: SRR for MMP and AC as a function of molecule number for a glockenspiel.

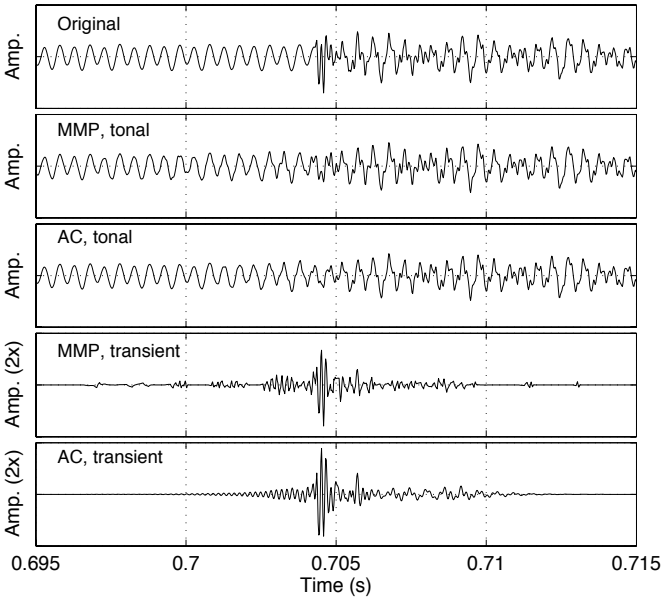

Figure 8: Glockenspiel segment approximated by tonal and transient molecules produced by MMP and AC.

Figure 6 illustrates the agglomeration process.

4 Simulation Results

The following settings were used in the computer simulations. For MMP, the MDCT uses a sine window of length $s = 1024$ samples, the Daubechies filter of length 4 is used in the wavelet transform, the maximum wavelet scale is $J = 9$ (512 samples), the order in (8) is $W = 5$, and $\epsilon = 0.02$. In AC, $\sigma = 600$, $\rho_{\min} = 0.01$, and $\mu_{\min} = 30$ ms. AC is performed with decompositions having SRR = 30 dB found by MP.

For the signal shown in Fig. 4, MMP extracts 98 molecules (93 tonal, 5 transient), representing the signal to SRR = 26.8 dB. From a sparse approximation of 976 Hann atoms, AC produces 71 molecules (62 tonal, 9 transient) of two or more atoms, representing the signal with an SRR = 23.19 dB. The growth of the SRR for each method as a function of molecule number is shown in Fig. 7. From (11), MMP extracts molecules in the order of their energy. The molecule order for AC is instead related to the signal time-line [3].

Segments of the resulting waveforms produced by each method are shown in Fig. 8. The greatest difference between the two methods can be seen in the transient molecules. Synthesizing the original signal using the molecules generated from each method produces similar sounding results, but the residuals of each sound

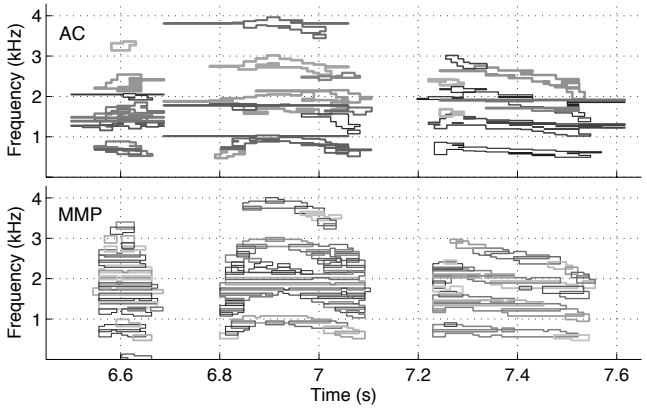

Figure 9: Segment of the bird call shown in Fig. 1 structured into tonal molecules by AC and MMP. Grayscale is used to offset molecules.

quite different, with MMP producing a "buzzy" residual. The transient molecules produced by MMP sound "crunchy," while those of AC are more impulsive. Both the transient and tonal signals of each method suffer from pre-echo. An extension to MMP to handle pre-echo is presented in [1], but was not implemented here.

Finally, returning to the example in Fig. 1, the resulting structured representations using MMP and AC are shown in Fig. 9, where the outline of the TF region of each molecule is shown. Observe that MMP cannot accommodate sweeping frequencies, which are broken into smaller units, whereas AC is able to handle these (see, e.g., the range 7.2-7.6 s). However, AC misses several significant portions of the signal, such as the onset at 6.8 s in the harmonics. Note also the narrowband "skewers" extending into TF regions that have no energy in the spectrogram in Fig. 1, e.g., the atom at about 4 kHz starting near 6.7 s. These are caused by the greedy atom selection of MP in (2) [11].

5 Conclusion

We have reviewed two methods for generating sparse and structured representations of acoustic signals: a joint approach (MMP) that builds structures as part of the decomposition process, and a sequential approach (AC) that builds structures from a sparse approximation using rules for clustering. We have also illustrated their properties using a musical and ecological signal.

With its choice of analysis dictionaries, MMP is remarkably fast when it employs a MDCT for finding and building tonal molecules, and a critically sampled multirate perfect reconstruction filterbank for the transient molecules. These choices, however, make the molecules translation variant. An example of this is clearly seen in Fig. 5 where the transient occurs in the latter half of the wavelet tree. This could cause a portion of a transient to be lost or split across wavelet trees. The definition of the local tonality index in (8) also causes problems when a signal has changing frequencies. We see this in Fig. 9 where such instances are split over several molecules.

On the other hand, because it relies on an OM, AC is much slower for building structured representations, and it inherits the problems of OMs [11], but one is free to use any dictionary. Furthermore, the similarity rules are flexible, and the general ones presented here are able to handle signals with changing frequency content. With a multiresolution dictionary providing a sufficiently redundant tiling of the TF plane, AC is less sensitive to translation than MMP.

The structured representations resulting from these methods can be used in several applications, including signal analysis, coding, content discrimination, and transformation. There is also the possibility of using these structured representations for generating sparse thumbnails of signal content for database indexing and querying. Future work will examine more complex rules for agglomeration, the effects of noise and polyphony, and a combination of the two methods such that AC is performed on the results of MMP.

Acknowledgments

The authors wish to thank Dr. Pierre Leveau. This work was supported in part by the National Science Foundation under Grant CCF 0729229.

References

[1] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1808–1816, Sept. 2006.

[2] P. Leveau, E. Vincent, G. Richard, and L. Daudet, "Instrument-specific harmonic atoms for mid-level music representation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 116–128, Jan. 2008.

[3] B. L. Sturm, J. J. Shynk, and S. Gauglitz, "Agglomerative clustering in sparse atomic decompositions of audio signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process.*, Las Vegas, NV, Apr. 2008, pp. 97–100.

[4] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, Aug. 1998.

[6] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic, Norwell, MA, 1991.

[7] M. Aharon, M. Elad, and A.M. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[8] J. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[9] M. Bosi and R. Goldberg, *Introduction to Digital Audio Coding and Standards*, Kluwer Academic, Boston, MA, 2003.

[10] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, CA, 2nd edition, 1999.

[11] B. L. Sturm, J. J. Shynk, L. Daudet, and C. Roads, "Dark energy in sparse atomic estimations," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 3, pp. 671–676, Mar. 2008.