A COMPARISON OF TWO EXTENSIONS OF THE MATCHING PURSUIT ALGORITHM FOR THE HARMONIC DECOMPOSITION OF SOUNDS

*Sacha Krstulović, Rémi Gribonval**

IRISA
Campus de Beaulieu – 35042 Rennes Cedex, France
{sacha,remi}@irisa.fr

*Pierre Leveau, Laurent Daudet**

Laboratoire d'Acoustique Musicale
Univ. P&M Curie/Paris 6 – 11 r. Lourmel – 75015 Paris, France
{leveau,daudet}@lam.jussieu.fr

ABSTRACT

In the framework of audio signal analysis, it is desired to obtain sparse representations that are able to reflect the harmonic structures, e.g., issued from musical instruments. In this paper, we compare two approaches which introduce some explicit models of harmonic features into the Matching Pursuit analysis framework. The first approach is the Harmonic Matching Pursuit (HMP), where the harmonic structures are modeled by sets of harmonically related Gabor atoms which are directly optimized in the analysis loop. The second approach, called Meta-Molecular Matching Pursuit (M3P), is based on the *a posteriori* agglomeration of elementary features coming from a Short Time Fourier Transform. We discuss the pros and cons of each method through experiments involving different audio signals, and conclude on possible approaches for combining the two methods.

1. INTRODUCTION

For the efficient representation of audio signals, it is hoped that sparse representations will better capture the salient features of the signals of interest. Musical sounds, for instance, can be characterized by the presence of harmonic structures, linked with the acoustics of the most common musical instruments.

The automatic extraction of harmonics or partials has been addressed by different approaches [1, 2, 3]. Most of the methods involve a peak-picking based on a Short Term Fourier Transform (STFT), then an *a posteriori* grouping of the harmonically related peaks. These approaches are constrained by the fixed time-frequency resolution of the STFT. Conversely, the Matching Pursuit algorithm implements a multi-resolution analysis by interpreting a signal as a sum of elementary *atoms* which do not need to be orthogonal. In the early versions of this algorithm, the elementary atoms were defined as Gabor Atoms, i.e. elementary sine waves with a modulation in amplitude and scale. But this specification is not intrinsic to the algorithm: its structure makes it possible to introduce atoms or molecules that correspond to signal features of a more elaborate nature than localized waveforms.

This paper compares two extensions of the Matching Pursuit algorithm which introduce an explicit modeling of harmonic structures for the analysis of audio signals. A general presentation of the Matching Pursuit algorithm will be given in section 2. A first variant, called Harmonic Matching Pursuit (HMP), defines some multi-scale harmonic molecules as groups of harmonically related Gabor atoms. These molecules can be optimized directly in the analysis loop [4]. This method will be presented in section 3. A second variant, called Meta-Molecular Matching Pursuit (M3P), begins with tracking the partials in a single-scale STFT representation of the signal [5], then groups them according to potential harmonic relations. This method will be presented in section 4. Pros and cons of each method will be discussed in section 5, and illustrated by the decomposition of different audio signals. Section 6 will conclude on possible approaches for combining the advantages of both methods.

*This joint work was partially supported by the MathSTIC program of the French CNRS.

2. THE MATCHING PURSUIT ALGORITHM

Matching Pursuit (MP) is part of a class of signal analysis algorithms known as Atomic Decompositions. These algorithms consider a signal $\mathbf{x}$ as a linear combination of known elementary pieces of signal, called atoms, chosen within a dictionary $\mathcal{D}$:

$$\mathbf{x} = \sum_{m=1}^{M} \alpha_m \mathbf{w}_m \qquad \text{where } \mathbf{w}_m \in \mathcal{D}. \tag{1}$$

Usually, the dictionary $\mathcal{D}$ is overcomplete: in dimension N, this means that $\mathcal{D}$ has more than N elements and spans the entire space. In this case, the above decomposition is not unique – there may even be an infinite number of solutions. Among all possible decompositions, the preferred ones are the compact (or "sparse") ones, which means that only the first few atoms in Eq. (1), in order of decreasing α_m, are needed to obtain a good approximation of the signal. In general, the bigger the dictionary, the greater the number of potential solutions, and thus the better the chance of finding a more compact signal approximation. However, for general overcomplete dictionaries, finding the truly optimal decomposition according to some pre-determined optimality and compactness criteria is a nontrivial task. As a matter of fact, this problem has received a lot of attention, with a number of methods aimed at finding an optimal approximation of the signal for specific optimality criteria [6, 7, 8]. Unfortunately, a heavy computational cost is usually associated with these methods, which prevents them from being practically applicable to large data files such as audio signals.

As an alternative to global optimization techniques, the Matching Pursuit algorithm (MP), originally introduced in [9], is a fast iterative method which tackles the problem by operating a local optimization. At each iteration m, the algorithm looks for the atom $\hat{\mathbf{w}}_m$ which is the most strongly correlated with the signal, i.e. which has the highest absolute scalar product with the signal $\mathbf{x}$:

$$\hat{\mathbf{w}}_m = \arg\max_{\mathbf{w}\in\mathcal{D}} |\langle \mathbf{x}, \mathbf{w} \rangle| \tag{2}$$

The corresponding weighted atom $\alpha_m \hat{\mathbf{w}}_m$ is then subtracted from $\mathbf{x}$, with $\alpha_m = \langle \mathbf{x}, \hat{\mathbf{w}}_m \rangle$, and the pursuit is iterated on the residual $\mathbf{r} = \mathbf{x} - \alpha_m \hat{\mathbf{w}}_m$. When the desired level of accuracy is reached (e.g., in terms of the number of extracted atoms or in terms of the energy ratio between the original signal and the residual), the iterations are stopped.

The MP algorithm was initially used [9] with dictionaries of multiscale time-frequency Gabor atoms of the form:

$$\mathbf{w}_{(s,u,\omega)}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t-u}{s}\right) e^{2i\pi\omega t} \tag{3}$$

with $w(t)$ a Gaussian window of unit energy. These atoms correspond to localized sinusoids of duration s, location u and frequency ω. For such dictionaries, MP runs in a fraction of the time needed for the global optimization techniques, and provides decompositions that are quasi-optimal (or even optimal under certain sparsity conditions [10]).

Clearly, the choice of the dictionary $\mathcal{D}$ is of prime importance, since its atoms should ideally be able to fit the typical elementary features of the analyzed signals. Gabor atoms are well adapted to the representation of the low level structures compatible with localized sinusoids. In this respect, the corresponding MP decompositions have proved useful as a tool for multi resolution time-frequency analysis. However, in the context of audio signal analysis, it would be desirable to capture some higher level structures, such as harmonic features.

3. HARMONIC MATCHING PURSUIT

The Harmonic Matching Pursuit algorithm [4] extends the principle of Gabor-based Matching Pursuit by introducing some harmonic molecules described as:

$$\sum_{k=1}^{K} c_k \mathbf{w}_{(s,u,\lambda_k\omega)}(t) \tag{4}$$

This feature model describes a combination of Gabor atoms $\mathbf{w}$ which center frequencies are linked by an approximate harmonic relation corresponding to $\lambda_k \approx k$, and which amplitudes are weighted by the coefficients c_k.

The resulting harmonic dictionary is quite large. However, thanks to its particular structure and its connection with the Gabor dictionary, it is possible to design an efficient implementation of the algorithm [4]. Overall, the principle remains iterative: starting from an initial residual $\mathbf{r}_0$ which is the analyzed signal $\mathbf{x}$, each iteration uses two steps:

1. selection of the harmonic atom – characterized by its parameters of duration s_m, location u_m, fundamental frequency ω_m and amplitude/phase for each of K partials – which represents the best match to the residual. This is done by maximizing the criterion:

$$C(s,u,\omega) := \sum_{k=1}^{K} \left| \langle \mathbf{r}_{m-1}, \mathbf{w}_{(s,u,\lambda_k\omega)} \rangle \right|^2 \tag{5}$$

2. removal of the best matched atom to obtain the new residual, using the coefficients $\alpha_{m,k}$ estimated as:

$$\alpha_{m,k} = \langle \mathbf{r}_{m-1}, \mathbf{w}_{(s_m,u_m,\lambda_k\omega_m)} \rangle \tag{6}$$

After M iterations, and considering K partials in the harmonic model, the signal is decomposed as:

$$\mathbf{x}(t) = \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{m,k} \mathbf{w}_{(s_m,u_m,\lambda_k\omega_m)}(t) + \mathbf{r}_M(t). \tag{7}$$

This decomposition can further be used for particular processing purposes, such as rebuilding the signal with a selection of atoms, or issuing a time-frequency representation.

4. META MOLECULAR MATCHING PURSUIT

The Meta-Molecular Matching Pursuit (M3P) is an extension of the Molecular Matching Pursuit (MMP) introduced in [5]. The main idea of the MMP is to track the local structures which appear in the Short Time Fourier Transform (STFT). As a matter of fact, it defines sound molecules as groups of neighboring STFT atoms. In other words, each step of MMP consists in selecting a group of Gabor atoms $\mathbf{w}_{s,u_k,\omega_k}(t)$ with s a fixed single scale and $(u_k, \omega_k) \in I_m$, to form a tonal molecule, corresponding to a partial (horizontal line in the time-frequency plane). Building on MMP, M3P allows to select harmonically related tonal molecules $I_{m,n}, (n \in [1..K])$, subsequently grouped in harmonic combs $\mathcal{M}_m$ (called meta-molecules). The major difference with HMP is that, at each iteration m, the duration and shape of a harmonic molecule is not defined a priori by the model: these parameters result *a posteriori* from the grouping of STFT atoms. From an algorithmic point of view, the molecules are not optimized through the direct maximization of a correlation criterion; instead, the M3P algorithm defines a method to group short time, single scale atoms in molecules $I_{m,n}$ that will fit the harmonic structures of the signal.

This algorithm proceeds in two steps: first, the tracking of an individual partial likely to belong to a harmonic structure, then the fitting of a harmonic comb around the detected partial.

4.1. Tracking of an individual partial

First, a starting point for the tracking is determined from a *harmonicity index*. This index takes the form of a spectral product defined as:

$$H(u,\omega) := \prod_{p=1}^{P} T\left(u, \frac{\omega}{p}\right) \tag{8}$$

where:

- $T(u,\omega) := \frac{1}{W_s} \int_u^{u+W_s} |S(t,\omega)| dt$ is a local time-averaging of the STFT modulus, with W_s a time persistence constant;
- p represents a frequency-dilatation ratio;
- P can be set to 2 or 3, depending on the desired emphasis on harmonic structures.

The time-frequency representation corresponding to $H(u,\omega)$ reinforces the harmonic structures: its maximum (u_H, ω_H) indicates the location of a partial related to a harmonic comb.

¿From this starting point, the partial is fitted by tracking backward and forward in time the maxima of the STFT modulus. First, the STFT starting point $(u_{\text{STFT}}, \omega_{\text{STFT}})$ is identified as follows:

$$(u_{\text{STFT}}, \omega_{\text{STFT}}) = \arg \max_{\substack{u \in [u_H, u_H + W_s] \\ \omega \in [\omega_H - \Delta\omega, \omega_H + \Delta\omega]}} |S(u,\omega)| \tag{9}$$

where $\Delta\omega$ is the maximum frequency variation allowed around a central frequency ω_H within a single molecule.

Starting from $(u_{\text{STFT}}, \omega_{\text{STFT}})$, local maxima of the STFT are tracked iteratively both in the forward and backward directions in time, within a frequency range around ω_H. The stopping condition is the combination of a static threshold with a criterion of maximum energy decrease (resp. increase). Once the molecule/partial is delimited in time, the frequency thickness is set to the main lobe size of a sinusoid in the STFT representation (e.g. 3 frequency bins for a STFT with a Hanning window). An additional atom is added when a frequency variation is detected, to account for a related widening of the main lobe.

4.2. Extrapolation of a harmonic comb

Once a basic partial I_m is built, a set of templates is formed by thickening or narrowing the molecule in the time-frequency plane, then replicating it at the related harmonic frequency levels. Several candidate templates can be generated, depending on the hypothesized position of the partial: it could be either the fundamental frequency, or any of the harmonics.

The best template is selected as the one with the maximum energy per molecule, unless several candidates have close scores for this criterion (over 0.6 times the maximum of the criterion). In this case, the elected candidate is the one with the lowest frequency, since the other ones are likely to be overtones. Once the best fitting template is found, the corresponding meta-molecule $\mathcal{M}_m$ is subtracted from the signal and the algorithm is iterated.

5. RESULTS

To illustrate the performances of the presented methods, a trumpet phrase and a piano phrase, both sampled at 44.1kHz, have been analyzed. Figure 1 and figure 2 compare the spectrogram, the HMP and the M3P representations in the time-frequency plane. The spectrogram displays the magnitude of the STFT according to the time-frequency location with a grey level going from white for small magnitudes to black for the largest ones. Similarly, as explained in [9, 4] the HMP time-frequency representations is obtained by adding time-frequency representations of all the Gabor atoms involved in the decomposition (7), using their coefficient $\alpha_{m,k}$ to determine the grey level. Atoms at a large scale s_m correspond to thin horizontal lines while shorter scale atoms appear more concentrated in time and more spread in frequency. Harmonic molecules, which are visible as families of horizontal lines at harmonically related frequencies, correspond to a single object in the HMP decomposition. In the M3P representation a rectangle is drawn around each meta-molecule $\mathcal{M}_m$, and each Gabor atom which belongs to it is displayed with the appropriate grey level. The parameters used to obtain these representations are described in the following section.

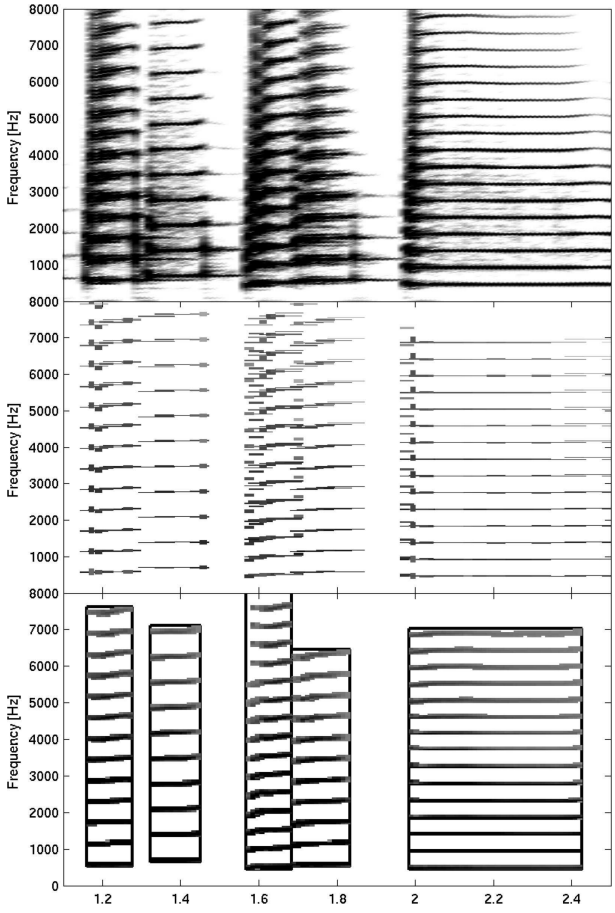

Figure 1: Spectrogram (top), HMP (middle) and M3P (bottom) representations of 5 trumpet notes.

5.1. Experimental setup

Spectrogram – The spectrograms are computed with 23ms (1024 samples) Hanning windows shifted by 6ms (256 samples).

HMP – The dictionary contains harmonic atoms corresponding to a range of 7 window lengths, logarithmically spread between 11.6ms (512 samples) and 92.8ms (4096 samples), all shifted by steps of 3ms (128 samples) and with a resolution of 4096 bins on the frequency axis. Gaussian windows are applied. The harmonic atoms are limited to $K = 15$ partials, searched above a fundamental frequency of 430Hz for the trumpet and 215Hz for the piano. The algorithm is iterated until the energy of the residual signal reaches -20dB below the energy of the original signal. This search yielded 102 harmonic molecules for the trumpet and 61 harmonic molecules for the piano. The trumpet signal lasts approximately 3 seconds, which corresponds to a search among 28 million atoms at each iteration. The piano signal lasts about 2 seconds, which corresponds to about 19 million atoms in the dictionary. With the fast implementation of MP developed at IRISA, completing these decompositions took about 1 minute on a Pentium 4@2.4GHz.

M3P – The dictionary is mono-resolution: STFT atoms are 23.2ms (1024 samples) long, and the overlap is set to half this length. The dictionary size is about 65000 STFT atoms per second of signal. A Hanning window is applied. Within each detected harmonic comb, the partials which energy is below 25 dB of the energy of the most energetic partial are rejected. On the piano sound, the proximity of the partials between successive notes has provoked partial jumps, leading to weak estimations of the harmonic structures. The search led to 10577 STFT atoms for the trumpet, grouped in 10 harmonic meta-molecules. It led to 4652 STFT atoms for the piano, grouped in 9 meta-molecules.

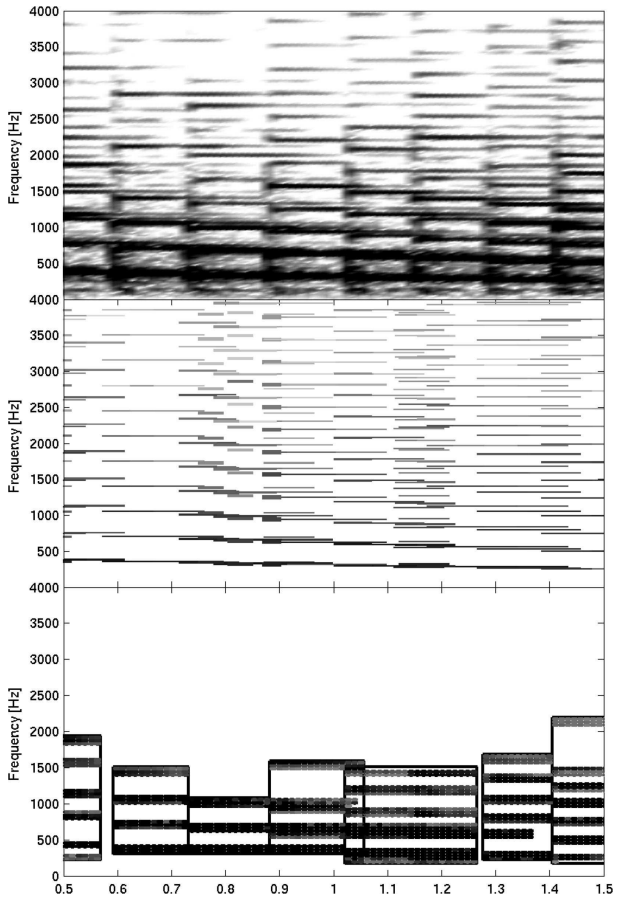

Figure 2: Spectrogram (top), HMP (middle) and M3P (bottom) representations of a range of piano notes.

5.2. Discussion of the results

Both methods are able to localize the harmonic structures with a reasonable accuracy. The HMP method uses a set of independent harmonic molecules which have various scales, but are bound to keep a constant frequency across their support. These molecules can be described with few parameters: HMP provides a sparse representation of the signal's harmonic phenomena. Nevertheless, the frequency modulated harmonic structures are fitted by sets of independent "flat" (non-chirped) harmonics which are not explicitly bound into a single object by the model.

Conversely, the M3P method builds its molecules by agglomerating (or "chaining") together many STFT atoms. The resulting meta-molecules provide an accurate description of the signal's harmonic phenomena in the frequency modulated cases. Nevertheless, the description of these meta-molecules invovles many parameters. Therefore, M3P provides a form of decomposition of an audio signal into structured objects, but the decomposition is not necessarily sparse.

It is important to notice that both methods are able to extract directly objects with a much longer time duration than an individual frame, where previous methods [1, 2, 3] perform harmonic grouping frame-by-frame. Also, it appears in the piano example that, although slight frequency variations from exact hamonic relations are allowed, some (more inharmonic) higher partials may be missed. Future improvement could include explicit inharmonicity laws, a straightforward extension in both HMP or M3P.

6. CONCLUSION

We have compared two techniques to decompose audio signals into hopefully meaningful elementary objects. HMP provides a multi-resolution analysis frontend which seems well adapted to signals composed of constant-frequency partials such as piano recordings. On such signals, it gives a meaningful decomposition into harmonic structures with very few harmonic molecules that seem to fit the notes. However, on sound signals with more frequency modulation such as bowed strings or trumpet, harmonic molecules are too "rigid" to represent what one could consider as elementary sound objects, which HMP decomposes into several pieces. In the latter cases, M3P seems flexible enough to represent frequency modulated harmonic objects as a single "meta-molecule" which chains together atoms admitting a single fine scale. The finer accuracy and increased flexibility of M3P for the decomposition of sounds into frequency-modulated objects comes however with a price, since a greater number of parameters is required to describe the objects (plus possible partial misses).

The experience gained by comparing the behaviour of both methods on various signals indicates that they are complementary, and it seems worthwhile combining them. We are currently investigating two possible approaches : some sound molecules could be defined as an agglomeration of multiscale HMP atoms, with a chaining method inspired from M3P; alternately, future developments of HMP could rely on a dictionary of chirped harmonic molecules instead of steady frequency ones.

7. REFERENCES

[1] X. Serra and J. Smith, "Spectral modeling synthesis:a sound analysis/synthesis based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[2] P. Depalle, G. García, and X. Rodet, "Tracking of partial for additive sound synthesis using hidden Markov models," in *Proc. IEEE ICASSP*, 1993, pp. 225–228.

[3] Y. Stylianou, "A pitch and maximum voiced frequency estimation technique adapted to harmonic models of speech," in *Proc. IEEE Nordic Sig. Proc. Symp.*, 1996.

[4] R. Gribonval and E. Bacry, "Harmonic decompositions of audio signals with matching pursuit," *IEEE Trans. Sig. Proc.*, vol. 51, no. 1, pp. 101–111, January 2003.

[5] L. Daudet, "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Trans. Speech and Audio Proc.*, 2005 (to appear).

[6] M. E. Davies and L. Daudet, "Sparse audio representations using the MCLT," *Signal Processing*, 2005 (to appear).

[7] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[8] I. F. Gorodnitsky and B. D. Rao, "Energy localization in reconstructions using FOCUSS: A recursive weighted norm minimization algorithm," *IEEE Trans. Sig. Proc.*, vol. 45, no. 3, 1997.

[9] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Sig. Proc.*, vol. 41, pp. 3397–3415, 1993.

[10] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Th.*, vol. 50, no. 10, Oct. 2004.