# AN EFFICIENT TWO-STAGE IMPLEMENTATION OF HARMONIC MATCHING PURSUIT

*Chris Duxbury, Nicolas Chétry, Mark Sandler and Mike Davies*

DSP & Multimedia Group
Department of Electronic Engineering
Queen Mary, University of London
Mile End Road, E1 4NS London, UK

chris.duxbury@elec.qmul.ac.uk, nicolas.chetry@elec.qmul.ac.uk

## ABSTRACT

We introduce an algorithm implementation for the decomposition of quasi-steady state audio signals using harmonic matching pursuits. Specifically, we propose an initial low-resolution pitch analysis followed by a high resolution harmonic grain extraction based on local complex interpolation within the spectral domain. We describe the implementation of the algorithm and illustrate its applications to musical analysis of monophonic signals before finally discussing possible improvements that could lead to the design of an iterative multi-pitch harmonic analysis system.

## 1. INTRODUCTION

Recent frameworks based on hybrid representation of audio signals attempt to decompose an input waveform as a linear combination of several distinct components. In such schemes, the stationary part is extracted from the original signal and separated from the noisy [1] or noisy plus transient [2] part. Resulting signals can then be processed individually, taking into account their own spectral, temporal or stochastic characteristics.

In this paper, we have focused on the modeling of the harmonic stationary component of audio signals. Such waveforms can be obtained by removing the transients [3], which occur during the attack of a note and present strong time localisation together with non-stationary properties. Diverse applications can be found in signal compression, musical analysis, musical modeling and automatic music transcription.

### 1.1 Harmonics, pitch and fundamental frequency

Many musical instruments produce harmonically related sounds, made up of several sinusoidal components, having their frequencies approximately integer multiples of the note's *fundamental frequency*. These sinusoidal components are also called *harmonics*, or *partials*, and give the sound its *timbral* characteristic, whilst the fundamental frequency generally gives the sound its *pitch*. However, in many cases, the fundamental frequency may not be present, without having an effect on the perceived pitch [4]. As an illustration, figure 1(a) shows a missing partial, and figure 1(b) shows a weak fundamental frequency component. These two sounds are however perceived as if the latter frequency components were unaltered.

Because of these issues, simply choosing the position of the lowest frequency high energy sinusoidal component as a possible fundamental frequency may not be relevant in terms of meaningful grain extraction.

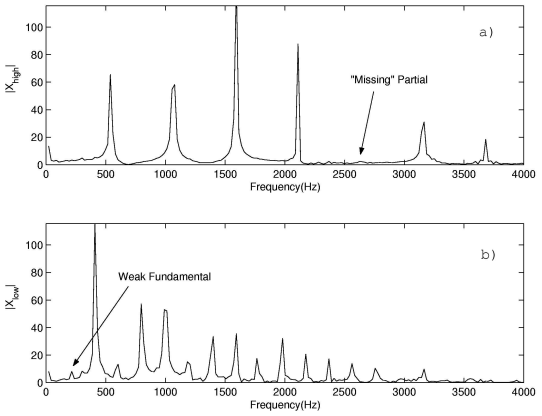


Figure 1: *Spectral plots of both (a) low and (b) high violin notes showing respectively a missing partial and a weak fundamental frequency.*

The low-resolution pitch analysis stage presented in section 2 helps to select the harmonic series which will remove the maximum energy from the spectra.

### 1.2 Matching pursuit using harmonic grains

Introduced in [5], the iterative harmonic matching pursuit algorithm approximates a signal with a linear combination of elementary waveforms, also called *harmonic atoms*. At each iteration, the best matching function is chosen from a redundant dictionary of harmonic Gabor functions. The residual is then similarly processed until the pre-defined energy-based stop criterion is satisfied.

Harmonic matching pursuit is an extension of the more general matching pursuit decomposition introduced in [6]. By considering dictionaries of harmonic atoms, such algorithms allow the extraction of higher level objects, which, in the case of audio signals, can lead to a meaningful musical interpretation (e.g note detection).

We present here an efficient dual resolution spectral-based approach that significantly reduces the computational requirements but still maximises the energy extracted at each stage:

- *Low resolution harmonic energy analysis*: the harmonic energy is calculated for each potential fundamental frequency. The harmonic series corresponding to the max-

imum harmonic energy is chosen. A rough value of the fundamental frequency is then estimated from the bin location of the fundamental and its partials.

- *High resolution harmonic grain extraction*: using the above value of the fundamental frequency, a local complex interpolation within the FFT frame is performed in order to determine more accurate values of the fundamental and its partials frequencies together with their corresponding amplitudes and phases. The resulting harmonic grain is then synthesised and subtracted from the original grain.

## 2. LOW-RESOLUTION HARMONIC ENERGY ANALYSIS

### 2.1 Harmonic energy calculation

Within an FFT frame, a frequency domain component contributes to the harmonic energy of the series if it is the maximum component within the corresponding *harmonic window*. This window increases in width linearly with the partial number. To be more specific, let us consider an FFT frame with a resolution of 20 Hz and a harmonic series corresponding to the fundamental frequency bin $100-120$ Hz. The first partial could appear anywhere between 200 and 240 Hz, corresponding to a *harmonic window* twice the width of the resolution value. Similarly, the next partial could appear anywhere between 300 and 360 Hz, i.e. three times the initial window width. The *harmonic window* width (in bins) as a function of the partial number $p$ is given by:

$$\nu(p) = p, \quad p = 1, ..., \rho$$

where $p$ is the partial number; $p = 1$ represents the fundamental and $\rho - 1$ the desired number of partials excluding the fundamental. Its value depends on the input signal and the sampling frequency. In practice, values around $\rho = 15$ for a 44.1 kHz sampled signal gave satisfactory results with our test signals.

The harmonic energy $\Lambda(k)$ is calculated for each potential fundamental index $k$ as:

$$\Lambda(k) = |X(k)|^2 + \sum_{p=2}^{\rho} \max_{\nu} |X(kp + \nu(p) - 1)|^2$$

where X is the complex FFT of the input frame. The harmonic series with the highest energy is then retained. It corresponds to the fundamental index $k = k_0$.

The selected series is used to calculate an approximation of the fundamental frequency:

$$\tilde{f_0} = \frac{1}{\rho} \sum_{p=1}^{\rho} \frac{f_{k_0,p}}{p}$$

where $f_{k_0,p}$ is the frequency corresponding to the maximum amplitude over the considered window $\nu(p)$:

$$f_{k_0,p} = f_{\max_\nu |X(k_0 p + \nu(p))|}$$

This new measure of fundamental is used to select the most relevant partial bins, by rounding the expected partial position

$$\tilde{f_p} = p\tilde{f_0}, \quad p = 1, ..., \rho$$

to the nearest bin value. These bins values are used in the high resolution harmonic grain extraction stage described in section 3.

## 3. HIGH-RESOLUTION GRAIN EXTRACTION

### 3.1 Frequency domain interpolation

Prior to the grain extraction, frequencies and corresponding phases of the selected fundamental and partials are interpolated in order to counterbalance the FFT finite resolution and therefore to maximise the energy extracted at each iteration.

Zero-padding the time domain input frame could achieve this goal but is clearly computationally inefficient as only the fundamental and the considered partials must be synthesised.

The following technique is identical to a zero-padding, but is applied locally to a region of the FFT around the considered peak (typically the width of the harmonic window as defined in section 2.1, extended by 2, 3 or 4 bins). It has the advantage of interpolating both phase and amplitude together.

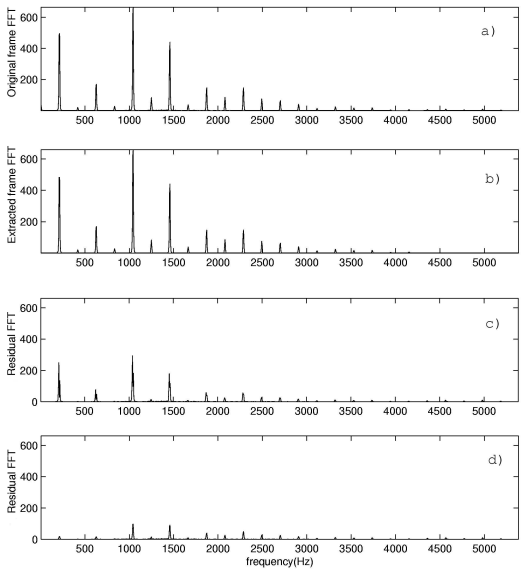


Figure 2: *Illustration of a harmonic grain extraction within a FFT frame. (a) original FFT frame. (b) FFT of the extracted grain after interpolation. (c) FFT of the residual without interpolation. (d) FFT of the residual after interpolation ($\xi = 20$).*

1. The interpolation function is calculated using the FFT of a zero padded rectangular window (by a factor $\xi$). These values are calculated once and saved in a table.
2. For each input frame, the frequency bins required for the interpolation (the fundamental and its related partials) are extracted as vectors and upsampled by the interpolation factor $\xi$.
3. The vectors are then convolved with the interpolation function values, achieving a local, complex domain interpolation.

Figures 2(c) and 2(d) show the advantages of using the interpolation in terms of the extracted energy per FFT frame.

The harmonic grain is then synthesised in the time domain and subtracted from the original grain. In the case of monophonic signals (one instrument), the algorithm stops. If not, the residual is processed similarly, until no more harmonic series are detected.

### 3.2 Orthogonality of harmonic grains

As shown in [5], the windowed sinusoids used in the algorithm are not necessarily orthogonal. Re-orthogonalisation, while possible, would be much too computationally expensive. In order to quantify the possible effects of using non-orthogonal basis functions, we compared the energy measures taken from the harmonic matching pursuits algorithm with the same decomposition using sinusoids which had been re-orthogonalised after selection using the QR orthogonal-triangular decomposition. The results for a violin tone are shown in figure 3. It can be seen that for the frame length considered here, the error due to non-orthogonality is trivial for all but the very low frequencies (in the case where the fundamental frequency is below 100 Hz).

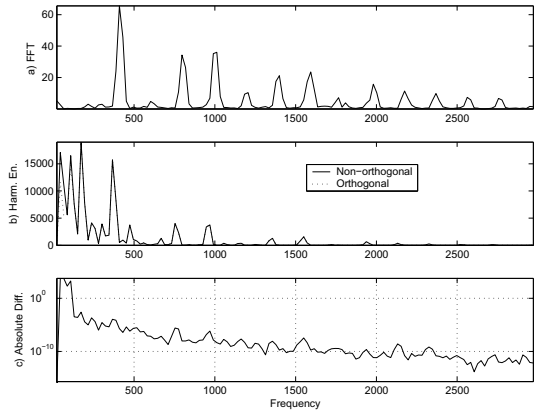


Figure 3: *Effects of orthogonalising sinusoidal components of a harmonic grain on the harmonic energy measure. The FFT is shown in (a), with the two different measures of harmonic energy given in (b) and the absolute difference between the two values in (c). Frequencies are in Hertz.*

### 3.3 Transient/Steady-State (TSS) separation

The harmonic matching pursuit algorithm is designed to model the slowly time varying steady state component of audio signals. Noisy high frequency time localised information, like transients or hard onsets, cannot be studied using harmonic grains. Further, their presences in the input signal may introduce artifacts as a *virtual* pitch (and possibly a whole harmonic series) can be detected during the low-resolution analysis stage. For this reason, a pre-treatment is applied in order to separate both transient and steady state signals from the original waveform. Details and examples of such *TSS* implementation can be found in [3].

## 4. ALGORITHM IMPLEMENTATION AND RESULTS

Tests audio signals (trumpet and jazz guitar pieces) were sampled at 44.1 kHz. The analysis stage was performed on successive frames of 2048 samples weighted by a Hanning window. Using such a window clearly sacrifices temporal localisation. A hop size of 25% (i.e. 512 samples) was therefore used to retain some signal timing information. $\xi$ was set to 20 and the local interpolation is applied on a 7 bins vector (the selected peak plus 3 bins on each side). A final point regarding robust implementation of this algorithm is to ignore the first three bins corresponding to the frequencies $0-65$ Hz, as issues related to the lack of orthogonality become more preponderant in that frequency range as stipulated in section 3.2. However, this should not induce severe artifacts as it is quite uncommon for musical signal to contain notes at such low frequency. Interpolated amplitudes, frequencies and phases are then used to synthesise the grain in the time domain using the oscillator bank approach (see [7] for details).

Figure 4 is an example of pitch extraction for a monophonic trumpet signal. The algorithm accurately restitutes the evolution of the fundamental frequency as a function of time. Figure 5 is an illustration of the complete application of the harmonic matching pursuits on a monophonic jazz guitar piece after having removed the transients [3]. Figure 5 (e) represents the time waveform residual after substraction of the synthesized signal from the original. One can notice that the transients are still slightly present, but nevertheless much more attenuated than if the whole original signal was used during the decomposition. Finally, figure 6 is an example of notes extraction. An artificial mixture of two piano notes without overlapping harmonics (B-250 Hz and F-350 Hz) has been synthesised. Two successive iterations of the algorithm are needed to decompose the input frame in two harmonic grains corresponding to the two individual notes. Residual errors (figures 6(c) and 6(f)) mainly contain noise and informal listening tests did not show any perceivable differences between the originals and extracted notes.

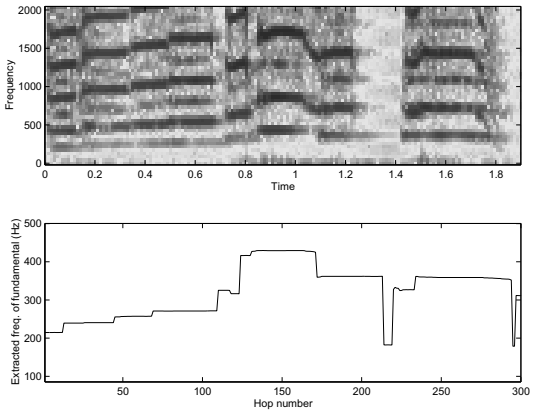


Figure 4: *Extracted pitches (fundamental frequencies) as a function of time for a harmonic monophonic trumpet piece.*

## 5. CONCLUSION AND FUTURE WORK

We introduced in this paper an efficient two-stage implementation of the matching pursuits algorithm based on a harmonic grain extraction within the spectral domain.

We first showed the advantages of considering harmonics objects in terms of:

- *high level musical interpretation:* as the extracted grain contains a complete harmonic series, it can be usefully characterised as a distinct musical note.
- *computational efficiency:* at each iteration, a complete series of sinusoidal components is extracted, as opposed to sinusoidal matching pursuit, where only one peak is picked at a time, thus requiring far fewer iterations.

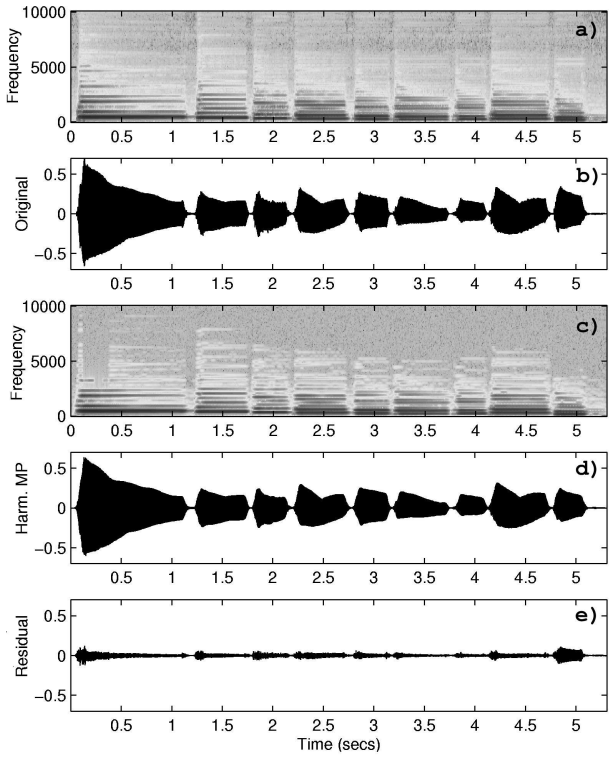


Figure 5: *Harmonic matching pursuits complete analysis/synthesis decomposition of a monophonic jazz guitar signal. (a) spectrogram of the original signal. (b) original waveform. (c) spectrogram of the resynthesised using a single harmonic grain per frame. (d) resynthesised waveform. (e) left over time domain residual.*

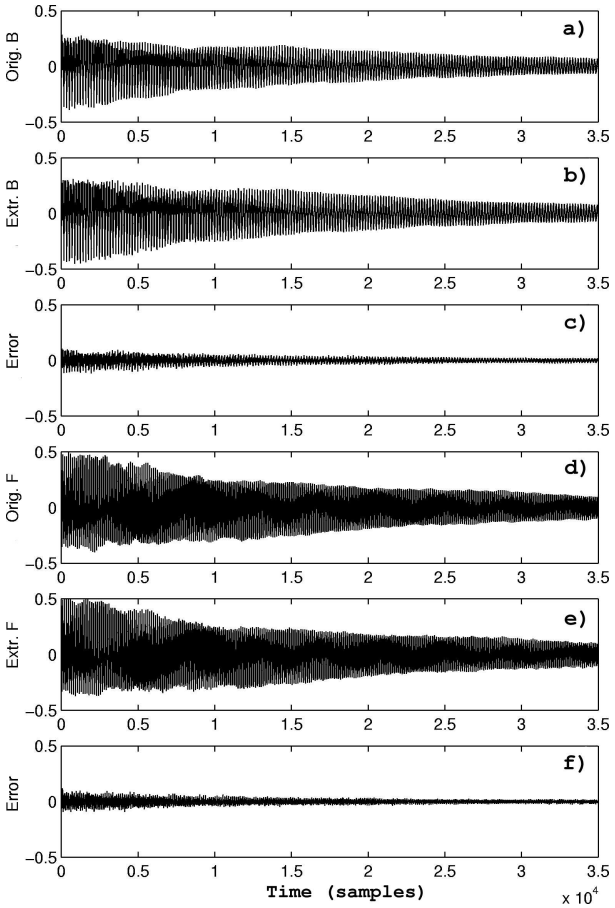


Figure 6: *Example of two piano notes separation. (a) and (d) are the original waveforms (respectively B-250Hz and F-350Hz), (b) and (e) the resynthesised signals, (c) and (f) the corresponding residual errors.*

Then, we introduce a local interpolation scheme, equivalent to a local zero-padding in order to obtain accurate frequency, amplitude and phase values of the selected harmonic peaks.

This implementation performs well for stationary audio signals. The quality of the extraction is good even though we make no assumptions about the type of instrument which is played.

Improvements, however, are needed to make the algorithm more robust and applicable to a wider range of signals. Firstly, problems arise with polyphonic audio mixtures containing overlapping harmonics. This is a common drawback of all the spectral analysis techniques. In such a case, harmonics corresponding to a given pitch may be affected to another harmonic series, thus introducing some false notes in the considered grain. This can be overcome for example by making assumptions about the harmonic distributions [8]. Secondly, if the signal is not purely steady state, the residual is shaped into harmonics which may introduce artifacts in the extracted signal. Particular attention should therefore be paid regarding the quality of the *TSS* decomposition.

## REFERENCES


[1] Xavier Serra, *Musical Signal Processing*, chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–123, Swets and Zeitlinger Publishers, 1997.

[2] L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, 2002.

[3] C. Duxbury, M. Davies, and M. Sandler, "Extraction of transient content in musical audio using multiresolution analysis techniques," in *Proc. Digital Audio Effects Conference (DAFX'01)*, 2001.

[4] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 1997.

[5] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, jan 2003.

[6] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionnaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec 1993.

[7] X. Amatriain, J. Bonada, and A. Loscos, *DAFX - Digital Audio Effects*, chapter Spectral Processing, pp. 373–438, Udo Zölzer, 2002.

[8] A. P. Klapuri, "Mutipitch estimation and sound separation by the spectral smoothness principle," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2001.