A review on techniques for the extraction of transients in musical signals

Laurent Daudet

Laboratoire d'Acoustique Musicale
Université Pierre et Marie Curie (Paris 6)
11 rue de Loumel, 75015 Paris, France
daudet@lam.jussieu.fr
http://www.lam.jussieu.fr

Abstract. This paper presents some techniques for the extraction of transient components from a musical signal. The absence of a unique definition of what a "transient" means for signals that are by essence non-stationary implies that a lot of methods can be used and sometimes lead to significantly different results. We have classified some amongst the most common methods according to the nature of their outputs. Preliminary comparative results suggest that, for sharp percussive transients, the results are roughly independent of the chosen method, but that for slower rising attacks - *e.g.* for bowed string or wind instruments - the choice of method is critical.

1 Introduction

A large number of recent signal processing techniques require a separate processing on two constitutive components of the signals : its "transients" and its "steady-state". This is particularly true for audio signals, by which we mean primarily music but also speech and some environmental sounds. Amongst all applications, let us mention : adaptive audio effects (enhancement of attacks [1], time-stretching[2], ...), parametric audio coding [3] (the transients and the steady-state are encoded separately), audio information retrieval (transients contains most of the rythmic information, as well as specific properties for timbre identification). Furthermore, transients are known to play an important role in the perception of music, and there is a need to define perceptually-relevant analysis parameters that characterize the transients.

However, there is a plethora of methods for the Transient / Steady-State (TSS) separation, with little indication of their relative merits. This arises from the fact that there is no clear and unambiguous definition of what a "transient" is, not what "steady-state" means for musical signals that are by essence non-stationary. In mathematical terms, this is an ill-posed problem, that can only lead to some tradeoffs. Indeed, we shall see that every definition leads to a specific decomposition scheme, and therefore different results in the separated TSS components. The goal of this paper is to make a review of some commonly used techniques, together with comparative results. Some of these methods are

recent developments, but others, that are well described in the literature, are mentioned here for the sake of completeness. Although we do not claim any sort of exhaustivity, we hope that we have covered the most important ones that have been successfully used for music. Obviously, a lot of other techniques could be used, that are not described in this article, since they were developed for other classes of signals (*e.g.* transient detection in duct flows, in underwater acoustics, in engine sounds). Our choice is to focus on musical signals, with a special emphasis on methods where the author had some hands-on experience, and where the computational complexity is reasonable so that they could be realistically applied to real full-band audio signals.

At this point, we should make it clear that the problem of TSS separation is related to, but distinct from, other classical musical signal processing tasks that are the classification of segments into transients or steady-state (for instance the way it is done in subband audio codecs such as MPEG 1 layer III, for the decision between the long and short window mode), or the binary TSS segmentation in time. On the contrary, all methods suggested here assume an *additive* model for the sounds, where transients and steady-state can in general exist simultaneously. TSS separation is also different from the problem of note onset detection [4], although it is clear that they share common methods.

The different methods can be grouped into 3 classes (even though this classification is certainly not unique), depending on the structure of their outputs (see table 1).

The fist class of methods, amongst the simplest in their principle, are based on linear prediction (section 2). They provide a decomposition of the sound into its excitation signal and a resonating filter. If the filter has been well estimated, most of the energy of the excitation signal is located at attack transients of signals.

The second class of methods (section 3) do not define transients directly, but rather extract from the signal its "tonal" part (also called sinusoidal part). If this extraction is successfully applied, the residual signal exhibits, as in the linear predictions methods above, large bursts of energy at attack transients. It also contains some slowly-varying stochastic residual.

Finally, the last class of signals (section 4) assume some explicit model for the transients, and the output of the model is 3 signals, that can be summed to reconstruct the original : one for the Sinusoidal part, one for the transients, one for the Residual (these models are often called STN models, for Sines + Transients + Noise).

Some results are presented in section 5, where some of the above approaches are compared on test signals. A tentative guide for the choice of a method most suitable for the problem at hand is finally presented, based on their pros and cons, that have to be balanced with computational complexity. The last section of this article (section 6) presents conclusions and future directions for research.

Table 1. Different transient extraction methods can be classified according to their outputs. For each class of methods, the signal related to the transients is highlighted

Method	Outputs			Section
Linear prediction	Resonance filter coefficients		**Excitation signal**	2
Tonal extraction	Tonal signal	**Non-tonal signal**		3
STN models	Tonal signal	**Transients signal**	Noise signal	4

2 Methods based on linear prediction

In this class of methods, the distinction between transient and steady-state is related to the notion of *predictability*. A steady-state portion of the signal is one where any part of this segment can be accurately predicted as soon as some small sub-sequence (the training sequence) is known.

Linear prediction in the time domain is a widely-used technique, since it is typically the core of most speech coding technologies. An the simplest auto-regressive (AR) case, the underlying idea is to consider the sound as the result of the convolution between an excitation signal and an all-pole filter. The Yule-Walker equations allow the estimation of the best order-P filter, that minimizes the energy of the prediction error. Once the filter is estimated, the excitation signal is simply the result of the filtering of the signal by the inverse filter (which only has zeros and therefore is stable). For steady-state parts of the sounds, the excitation signal can usually be simply modeled, for instance as an impulse train for voiced phonemes. More generally, this excitation signal will have a small local energy when the signal is highly predictable (steady-state portions), but energetic peaks when the audio signal is poorly modeled by the AR model. This typically corresponds to non-predictible situation, such as attack transients (or decay transients *e.g.* in case of a damper).

The obtained decompositions has a physical interpretation for source - filter models : when the excitation has a flat frequency response (impulse or whte noise), the AR filter is a good estimate of the instrument's filter. In the more general case when the excitation does not have a flat spectrum, it nevertheless provides a qualitative description of temporal and spectral properties of the sound, even though the physical interpretation is strictly lost.

Usually, this method gives good results when the signals are the result of some (non-stationary) excitation, filtered and amplified by a resonator. However, it has strong limitations : first, the order estimation (order of the filter) has to be known or estimated beforehand, which can be a hard task. Second, the estimation of the resonant filter will only be successful only if the training sequence does not contain a large portion of transients, and this makes the estimation on successive notes sometimes not reliable. However, this method is very well documented and easy to use in high-level DSP environments such as Matlab, and can be quite accurate on isolated notes.

Further extensions can be designed, for instance with auto-regressive with moving average (ARMA) models, but in this case the complexity is increased, both for the parameter estimation and for the inverse filtering that recovers the excitation signal.

3 Methods based on the extraction of the tonal content

In this class of methods, as in the linear prediction methods above, there is no explicit model for the transients. The goal here is to remove from the signal its so-called "tonal" or "sinusoidal" components. The residual is then assumed to contain mostly transients.

3.1 Segmentation of the Short-Time Fourier Transform

A natural way of expanding the above extraction methods is by using time-frequency analysis. The simplest implementation is the Short Time Fourier Transform (STFT), which provides a regularly-spaced local frequency analysis. Now, within each frequency subband, it is possible to perform the prediction search within each frequency band. The simplest model is based on the so-called "phase vocoder" [5], originally designed to encode speech signals. For the task of TSS separation, each time-frequency discrete bin will be labelled as "transient" (T) or "steady-state" (SS), and the underlying assumption is that we can neglect the influence of time-frequency bins that have a significant contribution in both domains. Note that we keep the terminology "steady-state" employed in the original papers, although the word "tonal" would be here more appropriate. After labeling, each signal, transient or steady-state, is reconstructed using only the corresponding time-frequency bins.

The simplest criteria for the identification of tonal bins is based on phase prediction in a given frequency bin k [5]. On steady-state portions of the sound, the (unwrapped) phase ϕ_n (n stands for the index of the time window) evolves linearly over time (hence the definition of instantaneous frequency as time derivative of the phase). Now, one looks at predicting the value of the phase ϕ_n in the current window, knowing its past values. A first-order predictor gives :

$$\phi_n^{pred} = 2\phi_{n-1} - \phi_{n-2} \tag{1}$$

Now, ϕ_n^{pred} is compared to the measured ϕ_n, and the labeling is based on the discrepancy between these two values :

$$\text{Time-frequency bin } (k,n) \text{ of type } \begin{cases} \text{SS} & \text{if } |\phi_n^{pred} - \phi_n| < \varepsilon \\ \text{T} & \text{otherwise} \end{cases} \tag{2}$$

where ε is a small constant that defines the tolerance in prediction error, for instance due to slight frequency changes, and it has to be adapted to the analysis hop size (number of samples between two analysis windows). Obviously, this

method can be seen as an extension of the linear prediction methods (section 2), as here a -basic- predictor is applied in every frequency bin.

More recently, this method has been refined in a few directions. It has been shown [2] that the results are significantly improved when processing the results in different subbands (with increasing time resolution at high frequencies), as well as by using an adaptive threshold for ε in equation 2. For onset detection purposes, the method has been further extended by using a complex-valued difference [6] that not only takes the phase difference into account, but also the magnitude (attack onsets are usually characterized by large jumps in amplitude).

3.2 Methods based on parametric representations : Sinusoidal models and refinements

Parametric representations assume a model for the signal, and the goal of the decompositions is to find the set of parameters that allow, at least approximately, to resynthesize the signal according to the model. For speech / music signals, it is natural to assume that the signals are mostly composed of tonal -*i.e.* sinusoidal- components, and here transients are defined as part of the non-tonal residual.

The simplest of this model was originally proposed by McAulay-Quatieri [7] for speech signals, where the sound is seen as a linear combination of a (relatively small) number J of sinusoids :

$$x(t) \approx \sum_{j=1}^{J} A_j(t) \sin\left(\varphi_j(t)\right) \tag{3}$$

where $\varphi_j(t) = \int_0^t \omega_j(\tau) d\tau + \varphi_j(0)$ represents the phase of the j-th partial. The parameters $(A_j, \omega_j, \varphi_j)$ for each partial sinusoid are assumed to evolve slowly over time, hence their values only have to be estimated frame-by-frame.

When applied to music signals, the residual contains all the components that do not fit into the model : stochastic components, or fast-varying transients. For general music processing purposes, this model has been refined to take into account the stochastic nature of the residual (Spectral Modeling Synthesis or SMS [8]).

3.3 Methods based on subspace projection

High resolution methods allow a precise estimation of exponentially damped sinusoidal components in complex signals. As in the linear prediction case, this model is physically motivated since exponentially-damped vibrations are the natural free response of oscillatory systems. The model is as follows :

$$x(t) = \sum_{j=1}^{K} A_j z_j^t + n(t) \tag{4}$$

where A_j is a complex amplitude, $z_j = e^{\delta_j + i2\pi f_j}$ is a complex pole that represents both the oscillation at frequency f_j but also the damping through the δ_j term,

and $n(t)$ is a noise term that is assumed white and gaussian. The principle of high resolution techniques is to estimate the values of all z_j through numerical optimization techniques. The obtained resolution is typically much higher than in the simple Fourier case. Specific methods have been proposed, that offer a better robustness to noise than the standard Pisarenko or Prony estimations methods. In particular, ESPRIT [9] and MUSIC [10] reduce this task to an eigenvector problem. When the number K of components is known, the span of the K obtained eigenvectors is called the "signal space", its complement is called the "noise space". Projecting the signal onto these 2 subspaces provides a natural decomposition of the signal into its tonal and non-tonal components. YAST [11], a recent variant of these methods for time-varying systems, is particularly suitable for music signals since it allows a fast processing of large-size signals. Note that the white noise hypothesis is generally not verified, in which case the processing has to be performed in separate subbands. Also, the number of components has to be known or estimated, and therefore the best estimates are obtained on isolated notes, where the number of components does not change over time.

4 Sines + Transients + Noise models

Three-components decompositions of the sounds have become very popular in audio coding, especially in the framework of MPEG-4. The aim is to decompose the file into three additive components, the tonal or sinusoidal component, the transients, and a slowly-varying wide-band stochastic component called "noise". The extraction can be done sequentially, first by a tonal component extraction as in section 3, and then by a transient processing on the non-tonal part. Alternatively, the separation can be done simultaneously, which usually gives better results but requires more computational power.

4.1 Sequential estimation of tones with hybrid dictionaries

In this TNS framework, the simplest method for TSS is to estimate each component at a time, first transient and then steady-state, or vice-versa. This is the basis for hybrid methods, that make use of two different orthogonal transforms. In the Transient Modeling Synthesis (TMS) scheme [3], the tonal part is first extracted by taking the large coefficients of a Modified Discrete Cosine Transform (MDCT). In a dual way, transients are analyzed in a pseudo-time domain constructed by taking the Fourier transform of the discrete cosine coefficients.

In [12], the tonal part is first estimated using the largest Modified Discrete Cosine Transform (MDCT) coefficients of the signal. Transients are then estimated by the largest Discrete Wavelet Transform (DWT) coefficients of the signal.

These methods are very simple, but suffer from two drawbacks: first, each component (T or SS) biases the estimate of the other component ; and second, at each stage the choice of the threshold between large "significant" coefficients,

and small "residual" coefficients is difficult. Although they may provide satisfactory results in some simple cases, the above limitations call for a simultaneous estimation of both TSS components, which is the topic of the next sections.

Simultaneous estimation by adaptive time-frequency resolution With adaptive time-frequency analysis, it is possible to obtain estimation of different components. the idea is to adapt *locally* the resolution of the transform to the signal. The choice of the resolution is based on some sparsity measure, for instance Shannon-like entropy measures. A simple version of this is the Best orthogonal Basis algorithm [13], where a multiresolution transform that has a tree-like structures (for instance, wavelet packets) adapts locally its resolution in time or frequency.

More recently, this idea has been extended and applied to music TSS separation, through "Time-frequency jigsaw puzzles" [14]. Here, this adaption step is made even more locally, in so-called *super-tiles* of the time-frequency plane. The algorithm is run iteratively until some convergence is reached. At every iteration, the algorithm finds, in each super-tile, the optimal resolution, transforms the signal accordingly, and subtracts the largest coefficients. As in many methods above, the choice of the threshold that governs, for a set of transform coefficients at a given resolution, what is "significant" and what is not is a critical point based mostly on empirical evidence.

4.2 Simultaneous estimation by sparse overcomplete methods

The goal of sparse overcomplete methods is to decopose the signal x as a linear combination of fixed elementary waves, called "atoms" :

$$x = \sum_k \alpha_k \varphi_k \ , \tag{5}$$

where α_k are scalars, and φ_k are the atoms drawn from a dictionary $\mathcal{D}$. In finite dimension, the dictionary $\mathcal{D}$ is said overcomplete when it spans the entire space and has more elements than the dimension N of the space. In this case, there is an infinity of decompositions of the form (5), and one would like to find one that is sufficiently sparse, in the sense that a small number $K \ll N$ of atoms provide a good approximation of the signal :

$$x \approx \sum_{j=1}^{K} \alpha_{k_j} \varphi_{k_j} \ , \tag{6}$$

If the dictionary is composed of two classes of atoms $\mathcal{D} = \mathcal{S} \cup \mathcal{T}$, where $\mathcal{S} = \{g_i\}$ is used to represent the tonal components of the sound (for instance long-window Gabor or Modified Discrete Cosine Transform atoms), and $\mathcal{T} = \{w_i\}$ is used to represent the transient part of the sound (for instance short-windows Gabor atoms, or wavelet atoms), a sparse approximation of the signal will provide a natural separation between transients and tones. In this case,

the noise is simply the approximation error, due to components that do not belong to either class, and the tonal layer (resp. the transient layer) is the partial reconstruction in the signal using only atoms in $\mathcal{S}$ (resp. atoms in $\mathcal{T}$).

However, for general overcomplete dictionaries, finding a good sparse approximation is a non-trivial task, and indeed it has been shown that finding the optimal K-terms approximation of the signal x is a NP-hard problem [15]. Many recent signal processing techniques have emerged recently (Basis Pursuit, Matching Pursuit, FOCUSS, ...), and we will here only a few of them that have been specifically applied to the TSS problem.

Matching Pursuit and extensions The Matching Pursuit [16] is an iterative method that selects one atom at a time. At every iteration, it selects the "best" atom φ_{k_0}, *i.e.* the one that is the most strongly correlated with the signal: $k_0 = \arg\max_k |\langle x, \varphi_k \rangle|$. The corresponding weighted atom is then subtracted from the signal $x \leftarrow x - \langle x, \varphi_{k_0} \rangle \varphi_{k_0}$ and the algorithm is iterated until some stopping criteria is reached (*e.g.* on the energy of the residual). This algorithm is suboptimal in the sense that, although it chooses at every iteration the atom that minimizes the residual energy, there is in general no guarantee that the set of selected atoms provide the best sparse approximation of the form (6). However, Matching Pursuit has become quite popular, mainly due to its simplicity, but also because experimental practice shows that in most cases the obtained decompositions are close to optimal (at least for the first few iterates).

For the sake of TSS separation, this algorithm has been extended into the Molecular Matching Pursuit [17], that at every iteration selects a whole group of neighboring atoms, called "molecule". The dictionary $\mathcal{D}$ is, as in the above hybrid model, the union of a MDCT basis (for tones) and a DWT basis (for transients), and a molecule is only composed of one type of atoms. Besides a reduced computational complexity, selecting molecules improves significantly the TSS separation over the original matching pursuit: first, it prevents isolated large atoms to be tagged as significant ; and second, it forces low-frequency large-scale components, that by themselves could equally go into transients or tones, into only one of these components according to the local context.

Global optimization techniques When the above results are not satisfactory, it may be desirable to use algorithms that choose a globally optimal or near-optimal solution, for a given optimality criteria. The problem can usually be written as a minimization problem:

$$u = \arg\min_u ||x - \Phi u||_2^2 + \lambda \|u\|_p^p \tag{7}$$

where Φ is the (rectangular) matrix of our overcomplete basis, and λ is a scaling parameter for the sparsity measure $\|u\|_p^p = \sum_{k=1}^M |u_k|^p$, with $0 < p \leq 1$.

The resolution of this problem typically involves very high computational costs. Many such techniques have been proposed in the literature, such as Basis Pursuit or FOCUSS. Amongst them, the Fast Iterated Reweighted SParsifier

(FIRSP) [18] has been successfully applied to audio TSS separation. The principle is to use the reweighted least squares algorithm for the optimization problem (7) with $p \to 0$ (this enforces a strong sparsity). For a practical implementation, this problem is reset in the Expectation Maximization (EM) framework. In general, this requires at every iteration the inversion of the matrix Φ, which can be very costly for large matrices.

However, this algorithm is of real practical use when the dictionary is the union of orthonormal bases, since in this case each EM iteration is replaced by a series of Expectation Conditional Maximizations within each orthonormal subspace, and the above matrix inversion reduces to a scalar shrinkage. Here, the strongly reduced computational complexity makes it a realistic choice for processing long audio segments. Further details can be found in [18, 19]. For TSS separation, the union of 4 orthonormal bases has been used, a long-window Modified Discrete Cosine Transform (MDCT), a long-window Modified Discrete Sine Transform (MDST), a short-window MDCT and a short-window MDST. After a few FIRSP iterations, the transients' energy is mostly concentrated on the short-windows transforms, and the SS in the long-windows transforms. Note that other orthogonal transforms can be used, for instance discrete wavelets for the transients ; however, preliminary results suggest that results are quite similar on most typical test signals.

Simultaneous estimation by adaptive time-frequency resolution With adaptive time-frequency analysis, it is possible to obtain estimation of different components. the idea is to adapt *locally* the resolution of the transform to the signal. The choice of the resolution is based on some sparsity measure, for instance Shannon-like entropy measures. A simple version of this is the Best orthogonal Basis algorithm [13], where a multiresolution transform that has a tree-like structures (for instance, wavelet packets) adapts locally its resolution in time or frequency.

More recently, this idea has been extended and applied to music TSS separation, through "Time-frequency jigsaw puzzles" [14]. Here, this adaption step is made even more locally, in so-called *super-tiles* of the time-frequency plane. The algorithm is run iteratively until some convergence is reached. At every iteration, the algorithm finds, in each super-tile, the optimal resolution, transforms the signal accordingly, and subtracts the largest coefficients. As in many methods above, the choice of the threshold that governs, for a set of transform coefficients at a given resolution, what is "significant" and what is not is a critical point based mostly on empirical evidence.

5 Comparative results

We have compared extraction results for 3 recent methods

- the YAST high-resolution method (paragraph 3.3). The signal is processed in 2500 Hz equal-width subbands, with 10 to 20 sinusoidal components par subband;

– the adaptive phase-vocoder approach (paragraph 3.1), described in [2];
– the jiigsaw puzzle approach (paragraph 4.2), in the variant TFJP2 described in [14].

Two soundfiles were tested, that have very different transient behavior: a glockenspiel, where the attacks are very sharp (the energy rising time is of a few ms); and a trumpet, where the energy rises on much longer time-scales (typically 50 ms, but this can extend to much higher values).

Separation results are presented in figures 1 and 2. On the glockenspiel example (figure 1), the results are quite similar for all three methods : the transient components exhibits sharp, high-amplitude peaks at note onsets. This is the typical case where all the definitions roughly agree on what a transient is.

On the contrary, the trumpet example exhibits very significant differences between the two methods. The YAST algorithm provides large energy bursts at the onset of notes, and smaller ones at their termination. The adaptive phase vocoder tends to capture more of the subtle variations within a note (see for instance the vibrato in the 6th note starting at about 1.5 s). Results of the jigsaw puzzle method are more difficult to interpret: even though its energy is well located on each onset, the amplitude of each onset transient varies somehow unexpectedly (see for instance the difference between the two notes starting at about 1s and 1.3 s). This lack of shift invariance is probably due to a particular choice of super-tiles.

It should be emphasized that these results are only indicative of the relative merits of each method. Results are highly signal-sensitive, at there is no guarantee that the performance of one algorithm will be similar on two signals that apparently belong to the same class. Furthermore, most of these methods requires a precise fine-tuning of the parameters, and for some of them the results are very sensitive to a particular choice of parameters.

However, we have tried in table 2 to summarize the main advantages and drawbacks of every method presented above. It should be emphasized that this comparison is only relevant within a category : linear prediction methods (described in section 2), Tonal extraction (described in section 3) or STN models (described in section 4). For each category, the balance between computational complexity and relevance of the results should help us in the choice of the most appropriate method for the problem at hand.

6 Conclusion

This paper is a first attempt to review and classify some techniques for the estimation of transients in music signals. Preliminary tests have been conducted; although a systematic comparative test is yet be performed, with more methods and more sound examples. However, a recurrent difficulty for such comparison is the lack of a common platform for testing : one of our medium-term goals is to develop such a software that could act as a unique front-end for some of the numerous methods above. Eventually, the main problem for the task

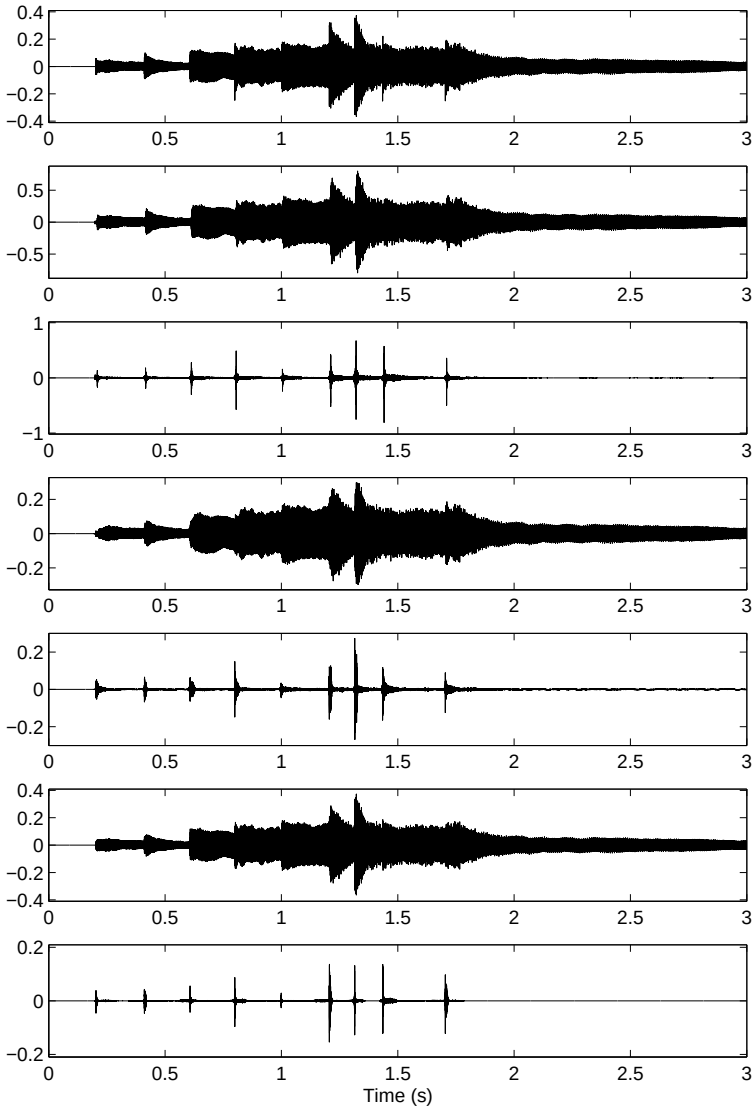

Fig. 1. Glockenspiel signal : comparison of three transients extraction techniques. From top to bottom:original signal, tonal part obtained by YAST, transients obtained by YAST, tonal part obtained by the adaptive phase-vocoder, transients obtained by the adaptive phase-vocoder, tonal part obtained by the jigsaw puzzles, transients obtained by the jigsaw puzzles.

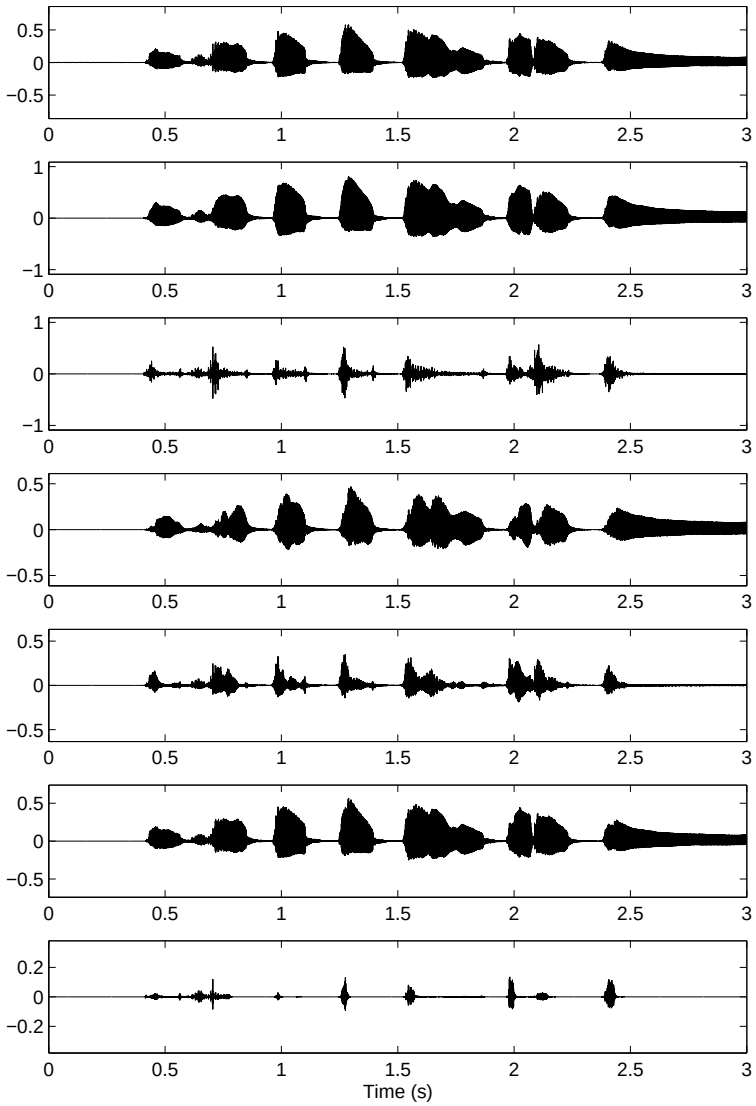

Fig. 2. Trumpet signal : comparison of three transients extraction techniques. From top to bottom: original signal, tonal part obtained by YAST, transients obtained by YAST, tonal part obtained by the adaptive phase-vocoder, transients obtained by the adaptive phase-vocoder, tonal part obtained by the jigsaw puzzles, transients obtained by the jigsaw puzzles.

Table 2. Tentative classification of relative computational complexity (from −−: very complex to ++: very fast), pros (P) and cons (C) for each method, as well as their most natural field of application.

Method	**Complexity**	**Pros, Cons & applications**
Linear prediction	+	P Source-filter interpretation C Relevant only for flat spectrum sources → *Physical models, lossless coding*
Model for sines only: transients in residual		
Adaptive phase-vocoder	+	P musically relevant C redundancy → *Audio effects, Preprocessing*
Sinusoidal model / SMS	+−	P Explicit signal model C No model for the residual → *Parametric coding, Audio effects*
Subspace methods	−	P High precision C Often requires hand-tuning → *Signal analysis, Preprocessing*
Model for sines and transients		
STN sequential estimation in orthonormal bases	++	P Fast algorithms C Difficult interpretation Threshold choices → *Transform coding*
STN simultaneous estimation by adapted fime-frequency tiles	+−	P General method C Threshold choices, no shift-invariance → *Analysis, Source separation ?*
STN simultaneous estimation by Matching Pursuit / Molecular MP	−	P Generates sparse data (and structured for MMP) C Optimality not guaranteed → *Parametric coding Source separation*
STN simultaneous estimation by global optimization (BP, FIRSP, . . .)	− to −−	P Very general, optimality criteria C Potentially very slow → *Analysis, Transform coding ?*

of comparing many TSS techniques is that one has to define one (or more) optimality criteria for deciding when a method is better than another. Some transientness criteria such as described in [1, 20] may be a first step towards relevant efficiency criteria.

One of our findings is that, unsurprisingly, the problem of TSS separation is indeed very different according to the nature of the signal. For sharp percussive sounds, the separation results are roughly independent of the chosen method -the simpler the better !-, but for slower rising attacks - *e.g.* for bowed string or wind instruments - the choice of method is critical. Finally, the biggest challenge is probably to link all these techniques to some perceptually relevant features, since numerous studies on music perception and timbre identification confirm the utmost importance of fast-varying transients. In the future, there is a need to develop a deeper understanding of the different time-scales involved in human perception. Finding perceptually-relevant signal parameters for transients is in our opinion one of the forthcoming challenges in the musical signal processing field.

Acknowledgements

The author wishes to thank Emmanuel Ravelli, Roland Badeau and Florent Jaillet for their help in the numerical examples. This work was supported by the French Ministry of Research and Technology, under contract ACI "Jeunes Chercheuses et Jeunes Chercheurs" number JC 9034.

References

1. Goodwin M. and Avendano C., "Enhancment of audio signals using transient detection and modification.," in *Proc. AES 117th Conv.*, San Francisco, CA, 2004.
2. Duxbury C., Davies M., and Sandler M., "Separation of transient information in musical audio using multiresolution analysis techniques," in *Proc. Digital Audio Effects (DAFx'01)*, Limerick, Ireland, 2001.
3. T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *Proc. of the International Computer Music Conference*, Greece, 1997.
4. Bello J.-P., Daudet L., Abdallah S., Duxbury C., Davies M., and Sandler M., "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, to appear.
5. Udo Zölzer, Ed., *DAFX - Digital Audio Effects*, John Wiley and Sons, 2002.
6. Bello J.P., Duxbury C., Davies M., and Sandler M., "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, 2004.
7. R.J. McAulay and Th.F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoust., Speech and Signal Proc.*, vol. 34, pp. 744–754, 1986.
8. X. Serra and J. O. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition.," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, winter 1990.

9. Roy R. and Kailath T., "ESPRIT - estimation of signal parameters via rotational invariance techniques," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 984–995, Jul. 1989.
10. Moon T. and Stirling W., *Mathematical Methods and Algorithms for Signal Processing*, Prentice-Hall, 2000.
11. Badeau R., David B., and Richard G., "Yet Another Subspace Tracker," in *Proc. International Conf. on Acoustics, Speech, and Signal Processing*, 2005, pp. 329–332.
12. L. Daudet and B. Torrésani, "Hybrid representations for audiophonic signal encoding," *Signal Processing*, vol. 82, no. 11, pp. 1595–1617, 2002, Special issue on Image and Video Coding Beyond Standards.
13. R.R. Coifman and M.V. Wickerhauser, "Entropy–based algorithms for best basis selection," *IEEE Trans. Information Theory*, vol. 38, pp. 1241–1243, March 1992.
14. Jaillet F. and Torrésani B., "Time-frequency jigsaw puzzle: Adaptive multiwindow and multilayered gabor expansions," *IEEE Transactions on Signal Processing*, submitted.
15. Davis G., *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, 1994.
16. S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, 1993.
17. Daudet L., "Sparse and structured decompositions of signals with the molecular matching pursuit," *IEEE Transactions on Speech and Audio Processing*, To appear.
18. Davies M. and Daudet L., "Fast sparse subband decomposition using FIRSP," in *Proceedings of the 12th EUropean SIgnal Processing COnference*, 2004.
19. Davies M. and Daudet L., "Sparse audio representations using the MCLT," *Signal Processing*, to appear.
20. Molla S. and Torrésani B., "Determining local transientness of audio signals," *IEEE Signal Processing Letters*, vol. 11, no. 7, July 2004.