AUTOMATIC SEGMENTATION OF THE TEMPORAL EVOLUTION OF ISOLATED ACOUSTIC MUSICAL INSTRUMENT SOUNDS USING SPECTRO-TEMPORAL CUES

Marcelo Freitas Caetano, Juan Jose Burred, Xavier Rodet

AUTOMATIC SEGMENTATION OF THE TEMPORAL EVOLUTION OF ISOLATED ACOUSTIC MUSICAL INSTRUMENT SOUNDS USING SPECTRO-TEMPORAL CUES

Marcelo Caetano, Juan José Burred [*]*, Xavier Rodet*

Analysis/Synthesis Team
IRCAM
{caetano,burred,rodet}@ircam.fr

ABSTRACT

The automatic segmentation of isolated musical instrument sounds according to the temporal evolution is not a trivial task. It requires a model capable of capturing regions such as the attack, decay, sustain and release accurately for many types of instruments with different modes of excitation. The traditional ADSR amplitude envelope model does not apply universally to acoustic musical instrument sounds with different excitation methods because it uses strictly amplitude information and supposes all sounds manifest the same temporal evolution. We present an automatic segmentation technique based on a more realistic model of the temporal evolution of many types of acoustic musical instruments that incorporates both temporal and spectro-temporal cues. The method allows a robust and more perceptually relevant automatic segmentation of the isolated sounds of many musical instruments that fit the model.

1. INTRODUCTION

The temporal evolution of musical instrument sounds plays a conspicuous role on the perception of their most important features [1]. Sound modeling and manipulation techniques could be greatly improved by the correct segmentation of musical instrument sounds taking into account the different characteristics of each perceptually different region. Notably, time stretching a sound ignoring its temporal evolution results in a perceptually different sound because the attack region, for example, is transformed by the same factor as the rest of the sound, even if it is well known [2] that the attack transients are perceived differently, being one of the most perceptually salient dimensions of musical timbre as unveiled by psycho-acoustical experiments [3]. The morphing of musical instrument sounds (especially those based on different modes of excitation) could largely benefit from a more accurate model of temporal evolution. A correct segmentation would also be beneficial to the modeling of musical instrument sounds for synthesis or classification purposes. Notably, the MPEG-7 standard [4] relies heavily on the estimation of perceptually important temporal events such as attack time to accomplish tasks such as automatic classification and similarity of sounds. Statistical and sinusoidal modeling could benefit from the segmentation as well. Computational [5] models of perceptual attack time (PAT) [6] must rely on an accurate estimation of attack time. However, to automatically estimate perceptually relevant features of musical sounds, such as the attack, from the signal we must rely on a model that utilizes features that exhibit a characteristic behavior during the events we wish to detect. The automatic segmentation task consists in the detection of events such as onset, attack, decay, sustain, release, and offset. There have been many proposals to the detection of some of these isolated events, especially onset [7], [8] and attack time [5]. However, the segmentation problem has been more rarely addressed [9], [10]. Usually, the problem is attacked in two steps: extraction of a detection function that supposedly contains information about the events to be detected followed by a technique responsible for automatically picking them out. The correct segmentation depends heavily on a model that accounts for as many different types of sounds as possible because not all instruments necessarily generate all the events. Also, due to the different nature of some of the events we wish to detect, it is almost naïve to expect one detection function to carry information about all of them. Historically, Helmholtz was the first to propose the segmentation of an acoustic musical instrument sound according to its temporal evolution [11], as early as 1885. This model led to the development of the attack-decay-sustain-release (ADSR) envelope [12], which was originally proposed for synthesis. Notably, it takes only amplitude information into account and supposes all sounds evolve the same way. Luce and Clark [13] suggested that the attack depends on the rise time of the amplitude envelope. Consequently, most detection techniques proposed use different estimations of the amplitude envelope to try and detect the events dictated by the ADSR model. Hajda [14] proposed a new model for the segmentation of isolated sustained (nonpercussive) musical instrument sounds that uses the relationship between the amplitude envelope and the temporal evolution of the spectral centroid to define the theoretical boundaries between perceptually salient segments of musical sounds. This model was coined the amplitude/centroid trajectory (ACT).

In this work, we propose to use the ACT model to automatically segment isolated acoustic musical instrument sounds. Our main goal is the automatic detection of the boundaries of the regions defined in the model for many types of instruments with different modes of excitation. For such, we will verify whether the model applies to some percussive instruments as well. An important part of our contribution lies in an improved technique to accurately estimate the amplitude envelope evolution of sounds by means of cepstral smoothing with a method known as *true envelope* [15], usually employed to estimate the spectral en velope. However, our main contribution lies in the automatic segmentation of the sounds by detecting significant changes in

[*] J. J. Burred is now with Audionamix, Paris, France.

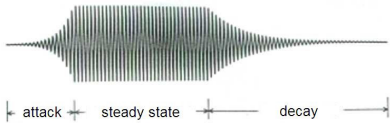

Figure 1: The Helmholtz model of the temporal evolution of acoustic musical instrument sounds. Helmholtz defined that the amplitude envelope can be divided into attack, steady state and decay. Figure from [17].

the features proposed in the ACT model, that only theoretically defines their boundaries. We compare our results with the baseline method proposed by Peeters [10]. The next section introduces the classical acoustic model of the temporal evolution of musical instrument sounds and the following section presents techniques to estimate the amplitude envelope and two segmentation techniques. We also introduce the improved true amplitude envelope estimation. Then, we proceed by a description of the signal level manifestation of the acoustic phenomena that generate the events we wish to detect. This exposition leads to the presentation of the improved ACT model. We finally present our technique to automatically detect the boundaries of the regions, followed by the conclusions and future perspectives.

2. THE CLASSICAL ACOUSTIC MODEL

Historically, Helmholtz was the first to propose the segmentation of isolated acoustic musical instrument sounds according to their temporal evolution [11]. Helmholtz characterized what he called musical tone as a waveform that follows an amplitude envelope that consists of the attack, the steady state and the decay, as shown in Figure 1. During the attack, the amplitude increases from zero to its peak value. In the steady state portion the amplitude is constant and finally decreases back to zero during the decay. Helmholtz concluded that sounds that evoke the sensation of pitch possess fixed waveforms that do not change in the course of the tone, apart from the amplitude envelope, whose temporal evolution has great impact on the perception of the tone, according to him. We should notice that this model only takes into account temporal cues provided by the amplitude envelope to define perceptually salient features such as the attack, steady state and decay.

Later on, Robert Moog devised the ADSR envelope model, shown in Figure 2, while developing his synthesizer and it quickly became the standard way to describe the amplitude envelope generator functions [12]. We should bear in mind that it was developed for synthesis purposes and, as such, it usually does not describe well the temporal evolution of most acoustical instrument sounds. However, most segmentation techniques [9], [10] rely on the detection of these events/regions based solely on the use of the amplitude envelope. Particularly, the attack is notoriously thought as being dependent on the rise time of the amplitude envelope [13] and some authors use it as its definition [8]. Therefore, we will present some amplitude envelope estimation techniques usually used in the detection of some of these events.

3. AMPLITUDE ENVELOPE ESTIMATION

The classic Helmholtz model led to the development of some segmentation techniques that only take temporal cues into ac count [9], [10]. Notably, these methods rely on the estimation of the amplitude (or energy, which is amplitude squared) envelope

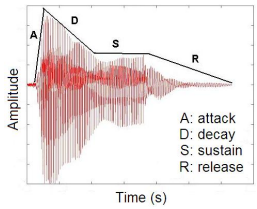

Figure 2: ADSR model applied to a wind instrument to explain its temporal evolution.

and use it as detection function to estimate the boundaries of the regions defined by the model. An early attempt [7] consisted of a piece-wise linear approximation of the waveform. The amplitude envelope is created by finding and connecting the peaks of the waveform in a window that moves through the data. Jensen [9] proposed a method that fits curve shape approximations to model the amplitude envelope of the partials of an additive model of instrument sounds and later Skowronek [18] applied it to approximate the global amplitude envelope. In this section we will compare the most widely used techniques to estimate the amplitude envelope and we will present a much more reliable amplitude envelope estimation method that optimally finds a smooth function that approximately matches the peaks of the waveform using true envelope cepstral smoothing [15]. The amplitude envelope estimation techniques that will be presented are the classical low-pass filtering, RMS energy, and analytic signal amplitude demodulation, as well as frequency-domain linear prediction [19] (FDLP) and our own proposal, true amplitude envelope (TAE). The aim of this section is to show that the TAE estimation technique leads to better results for all cases shown.

3.1. Low-Pass Filtering (LPF)

Low-pass filtering is the most straightforward way of obtaining a smooth signal that follows the amplitude evolution of the original waveform. It is based on a classical amplitude demodulation envelope follower technique [8], that low-pass filters a half-wave (*hwr*) or full-wave rectified (*fwr*) version of an amplitude modulated (AM) signal. The principle of amplitude modulation (AM) is that the amplitude changes of the signal carry the information we seek. There are many possible filter designs with different characteristics and the choice affects the quality of the final envelope. For instance, Jensen [9] proposes convolving the waveform with a Gaussian window function, resulting in a suboptimal estimation. Also, the cut-off frequency of the filter has a major impact on the result. High cut-off frequencies will likely produce an amplitude envelope with ripples and very low cut-off frequencies are less responsive to sudden amplitude changes.

3.2. RMS Energy

The RMS energy envelope is perhaps the most popular [20], [21], [22], [14] method for estimating the temporal evolution of the signal energy. The RMS energy envelope is based on the root mean square energy calculation and can be easily adapted to obtain an estimate of the amplitude envelope by simply applying it with a sliding rectangular window, as shown in equation (1)

$$RMS(t) = \sqrt{\frac{1}{T}\sum_{i=1}^{T} x_i^2(t)} \tag{1}$$

where $x_i(t)$ is the i^{th} local sample of the signal centered around t as seen through the window, t is the number of samples the analysis window moves, and T is the window length. The RMS is

a special case of the generalized mean with exponent $p=2$ and as such, also functions as a sort of moving average, low-pass filter that smoothes out the signal. The analysis step t imposes a trade-off between the temporal sampling rate of the envelope and how much information it represents. Small values of t react sooner to sudden changes in amplitude, while presenting ripple in more steady regions and larger values smooth out the ripples but tend to lag behind abrupt energy changes.

3.3. Analytic Signal

The Hilbert transform is part of a signal processing technique for amplitude demodulation [16]. The Hilbert transform of a signal $x(t)$ is defined as

$$\hat{x}(t) = x(t) * \frac{1}{\pi t} \tag{2}$$

where * stands for convolution. Using equation (2), we can define the analytic signal $z(t)$ as

$$z(t) = x(t) + j\hat{x}(t) = r(t)\exp[j\theta(t)] \tag{3}$$

The analytic signal is useful for envelope detection since its modulus $r(t)$ and time derivative of the phase $\theta(t)$ can serve as estimates for the amplitude envelope and instantaneous frequency of $x(t)$ under certain conditions. Notably, if the Hilbert transform of $x(t)$ is equal to its quadrature signal [16], then the estimates are equal to the actual information signals [16]. Synthetic (i.e., AM) signals can be constructed to have this property, but there is no reason to expect that acoustic musical instrument sounds also present it. A more realistic condition is verified when we are dealing with narrowband signals, which is rarely the case for musical instrument sounds. The Hilbert transform can be effectively used to extract the amplitude envelope of individual partials if applied to each frequency bin of the STFT, but when applied to the whole signal it is equivalent to trying to demodulate several AM signals at the same time.

3.4. Frequency-Domain Linear Prediction (FDLP)

Traditional linear prediction [23] estimates the spectral envelope from the time-domain signal. The idea behind FDLP [19] is to exploit time-frequency duality to extract the temporal amplitude envelope by applying linear prediction to a spectral representation. In particular, the used spectral representation is the discrete cosine transform (DCT), since it is real-valued. The envelope peaks, whose number and width are determined by the model order, will now be their frequency domain counterparts, the rectified waveform peaks. Thus, the model order has to be adjusted with respect to the temporal structure of the signal, and not to the formant structure of the spectrum.

3.5. True Amplitude Envelope (TAE)

True envelope [15] is a method for spectral envelope estimation that has shown better performance than linear prediction or cepstral methods such as discrete cepstrum, both in terms of accuracy and ease of model order selection. It is based on iteratively estimating the cepstrum and adapting it in such a way that the peak matching is maximized and inter-peak valleys are avoided. Here, we propose to use a dual of true envelope in the time domain. The time domain signal is subjected to the algorithm in stead of the Fourier spectrum. In this way, the amplitude envelope is expected to match the amplitude peaks more closely than the previously introduced methods. It is important to note that

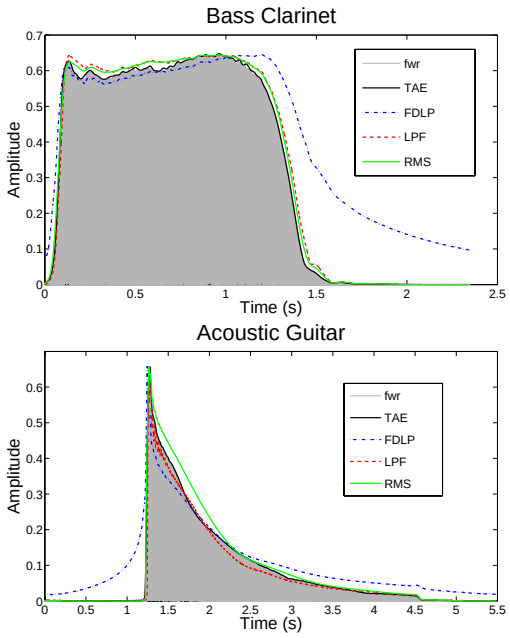

Figure 3: Full-wave rectified (*fwr*) waveform and amplitude envelope estimation methods; true amplitude envelope (TAE), frequency-domain linear prediction (FDLP), root-mean square (RMS) and low-pass filtering (LPF).

the optimal model order will be directly proportional to the fundamental frequency, rather than inversely proportional as in spectral envelope estimation.

Figure 3 shows a full-wave rectified (*fwr*) version of a bass clarinet and acoustic guitar waveforms and the amplitude envelope estimates. We are looking for the estimation that best fits the model waveform, that is, that follows the amplitude evolution most closely matching the peaks. It should be noted that all curves, except TAE, were normalized and scaled to the maximum of the *fwr* waveform and that RMS was also low-pass filtered to eliminate the ripples during mostly the sustain part. Upon close inspection, Figure 3 shows that TAE renders the best fit, closely following the peaks without ripples.

3.6. Segmentation Techniques Using the Amplitude Envelope

Finally, we present two previously proposed techniques to automatically segment individual musical instrument sounds based solely on the amplitude envelopes, namely derivatives [9], [18] and efforts [10]. Both methods try to detect the inflection points of the amplitude envelope based on the assumption that the amplitude envelope changes correspond to the boundaries of the regions we are looking for. Notably, these models define the attack as the rise time of the amplitude envelope, like other authors [8], [13]. Skowronek [18] proposed a segmentation method based on their attack-decay-and-sustain-release (A-D-&-S-R) model. They obtain an approximation of the amplitude envelope and use it together with its first derivative to estimate the boundaries of the three regions defined as start of attack (*soa*) and end of attack (*eoa*); and start of release (*sor*) and end of release (*eor*), as exemplified on the left of Figure 4. The right of Figure 4 shows the method of efforts, introduced by Peeters [10]. The method of efforts segments musical instrument sounds using the attack/rest model, whereby we separate the attack portion from the rest of the sound, be it sustained or percussive. First we divide the slope

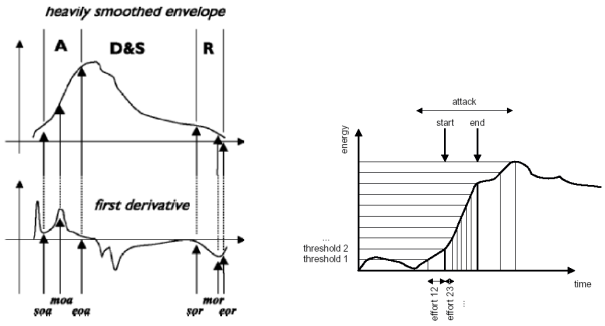

Figure 4: Depiction of the method of derivatives (on the left) and efforts (on the right) for the automatic segmentation. Figures respectively from [18] and [10].

corresponding to the rise time into N equal intervals according to the amplitude incremental values (called *thresholds*). We could define the start and end of the attack as fixed percentages of the minimum and maximum amplitude values, very much like Luce and Clark [13]. Instead, in this method, the start of attack and end of attack are estimated according to the slope of the rise region. So we calculate a piecewise measure of the slope for each threshold jump by measuring how long it takes to go from one threshold to the next (called *efforts*). The selected threshold is the one whose value is smaller than M times the mean threshold for both the start and end of the attack. They recommend to use M = 3 [10]. We will show in the next section why methods that rely solely on the amplitude fail to segment sounds into perceptually meaningful events because they use restricted information.

4. TEMPORAL EVOLUTION

The segmentation of musical instrument sounds depends on the correct detection of the boundaries of the regions. Clearly we need a good definition of the regions to be detected in order to be able to estimate them. The first problem we face is that not all instruments contain the same temporal events, so we cannot expect, for example, to be able to estimate the sustain part for a percussive instrument sound. This is where a robust model plays a key role in defining the segments and their boundaries. The most important aspect to be taken into account is a clear separation of cause and effect. The amplitude envelope is merely the description of one of the results of the source-filter interaction. It is fruitless to attempt to detect the boundaries of the events we want to estimate without a proper causal description. We must find the signal level counterparts of the physical events to properly estimate them. The technique we present in Section 5 uses spectro-temporal cues at the signal level left by the physical/gestural events to correctly segment them. The difficulty in this approach is that each instrument has its own particularities. Ideally, we search for a model that is robust enough to describe the signal level manifestations of as many types of instruments as possible. We will begin by describing the general model we will consider, namely, the source filter model, originally developed for speech; and then the physical characteristics of the events we aim to describe, and then later in Section 5 we will present the ACT model [14] for specific classes of instruments. Finally, Section 6 shows how we use the ACT model to obtain significant estimates of the boundaries from spectro-temporal traces left by the physical gestures.

4.1. Source-Filter Model

Listeners use many acoustical properties to identify sonic events, such as the spectral shape, formant frequencies, attack and/or onset and decay and/or offset, noise, among others [1]. The cues to identification and timbre vary across notes, durations, intensities and tempos. One model of sound production is based on two possibly interactive components, the source and the filter. The basic notion is that the source applies excitation energy to generate a vibration pattern composed of several vibration modes (modeled as sinusoidal components). This pattern is imposed on the filter, which acts to modify the relative amplitudes of the components of the source input. We obtain estimates of the source and the filter by calculating the spectral envelope, which is a smooth curve that approximately matches the peaks of the spectrum. The peaks of the spectral envelope (also called formants in voice research) correspond roughly to the vibration modes of the source-filter model. The number and absolute position of spectral peaks in frequency is important for musical instrument (source) identification. The relationship between fundamental frequency and timbre is readily apparent in some acoustic instruments. The clarinet, for instance, has three distinct registers, that is, three distinct pitch ranges with three different timbral characteristics. It is remarkable that a single instrument can have such a variety of timbres, but the example of the clarinet proves the impact of a resonating body on an instrument's timbre and temporal evolution. Resonators, by their nature, tend to amplify certain frequencies louder than others. These resonant frequency regions, called formants, are uniquely related to the size and shape of the instrument and its resonator. The relationship between applied energy and timbre is relatively clear. As more energy is input to the instrument, higher modes of vibration are achieved such that more partials are present in the frequency spectrum. This is why a note played *forte* is not just louder than *piano*, but also brighter in timbre.

In this work, we are primarily interested in the temporal evolution of the sounds, so we are going to examine the source-filter model from a temporal perspective. Figure 5 shows a simplified schematic view of the temporal evolution of the excitation (dotted line) and the resulting amplitude envelope followed by the sound (solid line) for two markedly distinct classes of excitation

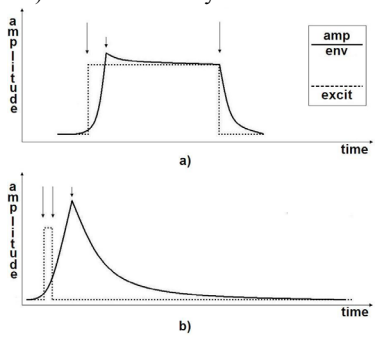

Figure 5: Simplified temporal evolution of the excitation (dotted line) and amplitude envelope (solid line) for two distinct classes of excitation modes. In a) we see the typical excitation and amplitude envelope resulting from the step-like excitation (e.g., blown/bowed) and in b) for impulse-like excitation (e.g., plucked/struck). The beginning and end of the excitation are marked with long arrows, while the short arrow shows the maximum amplitude attained by the resulting amplitude envelope.

methods, namely step-like (part a) and impulse-like (part b). Step-like excitation corresponds to playing modes whose energy is supplied for some length of time before being interrupted, while impulse-like is when energy is supplied in a short burst. The former typically applies to sustained sounds resulting from bowed strings and blown pipes, and the latter to percussive excitations such as plucked strings or struck instruments, although blown or bowed *staccato* notes would probably be better described by it. The beginning and end of the energy supply for each excitation mode are highlighted by long arrows and the short arrow marks the maximum amplitude attained by the sound. Next, we try to make a connection between the mainly physical events such as onset, attack, decay, sustain, release and offset and its model counterparts in connection with the excitation and resulting temporal evolution presented earlier. The idea is to find signal level manifestations of the physical gestures. Our main goal is to show that these events cannot be solely described by the amplitude envelope of most sounds, such that we need a more complete model to appropriately segment them.

4.2. Signal-Level Manifestation of Physical Events

Musical instruments are mechanical systems that by themselves are at equilibrium. They need an external source of energy input to produce sound. In general terms, all acoustic musical instruments have one (or more) method for applying mechanical energy to the system, herein termed the excitation method. Pianos have keys connected to hammers that strike a set of tuned strings. Violins have strings that are bowed or plucked. Clarinets have a mouthpiece with a single reed that, when blown, creates a vibrating column of air. Different modes of excitation will generally lead to perceptually different attacks.

The connection between intensity (dynamics), frequency and amplitude envelope is far less obvious. Before the onset, the instrument is in a state of equilibrium. As with all mechanical systems, there is a certain amount of resistance or inertia that keeps the instrument from vibrating on its own. Performers must overcome that inertia before their instrument will sound properly. The more energy a performer uses, the faster the resistance is overcome and the faster the instrument reaches its steady state vibration. For example, different fingerings on a wind instrument produce different lengths of air columns – longer columns mean more mass to vibrate. We know that large masses have more inertia to overcome, but also have more momentum once they are in motion. Thus, low notes have a longer attack and a longer release than high ones. In this section we define the physical/gestural events that generate/define each perceptually different region of the temporal evolution of musical instrument sounds and, more importantly, we find signal-level manifestations of the physical events.

4.2.1. Attack

The attack is perhaps the only event that is present in all sounds independent of the mode of excitation. The attack corresponds to the initial excitation of the instrument. The beginning of the attack is perhaps best characterized by the transition between no event and event (or more properly background noise and event for recordings, i.e., signals). This is usually termed onset. The end of the attack is more difficult to define since it depends on the physical gesture. Notably, transients occur until a permanent resonance mode is attained. For some instruments, we can make a clear distinction between the end of the attack and the beginning of the resonance. The time period when a hammer touches the piano strings would be the attack and the moment the standing wave pattern establishes itself in the string marks the beginning of the resonance mode. For a bowed string it is similar. From the moment when the bow first touches the string (onset) until the string enters a resonance regime with the bow we can devise two physically and perceptually distinct events. The end of the attack happens before the resonance. For a tube (blown instruments) the situation is similar.

4.2.2. Decay

The decay supposedly corresponds to a decrease of energy after the attack during which the permanent excitation regime (such as a standing wave vibration pattern) is already established. This is the region where the amplitude evolution of a percussive instrument sound constantly decays due to losses and strays from constant (step-like) excitation patterns (blown/bowed strings), where energy is repeatedly input to the instrument during a period of time. When we look closely, the decay remarkably contains standing wave patterns, even though the amplitude is decreasing. In plucked strings, for example, there is a clear spectral pattern that remains constant throughout and that is perceptually important. The decay contrasts with the amplitude evolution of a blown or bowed instrument, whose standing wave vibration pattern coincides with a more or less constant amplitude.

4.2.3. Sustain

The sustain part usually corresponds to the region where the system (musical instrument) is constantly exited with external energy. It is usually defined in terms of approximately constant amplitude. Perceptually, though, it is not reasonable to expect the region where a standing wave vibration pattern manifests as spectrally constant resonances (similar spectral shape) to be sufficiently described solely by the amplitude. Therefore we suggest that constant excitation instruments (bowed and blown, among others), where the energy and the spectral information remain roughly constant, present a sustained part.

4.2.4. Release

The release phase admits several interpretations, and its definition has not been consistent among authors. On the one hand, it can refer to the release of the excitation, such that the release segment is the interval between the time instant where the energy ceases to be supplied and the vibrations dying out (offset). This definition is common for sustained sounds (step-like excitation), but not always used for non-sustained sounds (impulse-like excitation), because in the latter case release would be equivalent to decay. On the other hand, it can refer to an intentional interruption of the vibration by the player. Most notably, in stringed keyboard instruments, this corresponds to the release of a key, which causes the damper to stop string vibrations. To avoid confusion between these very different physical events, we will use the following conventions. Release will correspond to the release of excitation in sustained instruments. Then, we will introduce a new segment called *interruption* to account for the case of intentional interruption of vibrations in non-sustained instruments.[2]

[2] Note that, while rare, it is also possible for the release phase of a sustained instrument to be followed by an interruption phase, such as when a violinist intentionally interrupts the vibrations of the strings after having stopped supplying energy to them by bowing.

So we see that although these regions share common spectro-temporal aspects, they cannot be defined simply in terms of energy evolution. Also, the accuracy of estimation depends on a detection function that captures the essence of the model and exhibits peculiar behavior at the boundaries of the regions that allow for a robust automatic detection.

5. AMPLITUDE/CENTROID TRAJECTORY MODEL

The classical Helmholtz model breaks down when we examine musical instrument sounds on a small scale. When the harmonic content of sound is examined with the STFT over small time periods, we discover that, contrary to the Helmholtz model, a sound's spectrum changes profoundly over time. During the attack portion of a sound, harmonic content may change rapidly and unpredictably. This phenomenon is called the initial transient. During the release, upper partials tend to disappear more quickly before the entire sounds fades away. While the sustain portion of the sound, when it exists, is certainly more stable than the attack or decay, it is hardly as static as Helmholtz would suggest. Clearly, the basic premise of the classical Helmholtz model - a static spectral envelope with a fixed amplitude envelope temporal evolution – is by no means an accurate and robust characterization of a wide range of acoustic musical instrument sounds. All these facts suggest that, in order to better understand the temporal evolution of sounds, we need a model that accounts for spectro-temporal changes. The vast majority of research in sound perception has focused either on the acoustic properties of musical instruments [24] or on the perception of sounds as unveiled by psycho-acoustic experiments [3]. The challenge we face today is to find the link between the two in order to be able to manipulate the sounds in a more perceptually meaningful way. A classical example is Risset's discovery that brassy trumpet sounds present a broader spectrum. The spectral centroid, defined in equation (4) reflects spectro-temporal acoustic properties at the signal level. Therefore, brassy sounds can be characterized as presenting a higher centroid value.

$$C(t) = \frac{\sum_{b=1}^{M} f_b(t) a_b(t)}{\sum_{b=1}^{M} a_b(t)} \tag{4}$$

Here $C(t)$ is the time-varying centroid, $f_b(t)$ is the frequency in Hz and $a_b(t)$ is the amplitude of frequency band b up to the M^{th} band computed.

Hajda [14] proposed a segmentation model he dubbed the amplitude/centroid trajectory (ACT) that relies on both the amplitude envelope and temporal evolution of the spectral centroid. In this model, the spectral centroid gives information about the excitation indirectly. The sudden transition characteristic of the onset reflects as a brief broadening and narrowing of the spectrum, causing the centroid to drop until the steady state resonance establishes itself, bringing the centroid up again to a somewhat steady value. For step-like excitation, the release is the moment when the player stops supplying energy to the instrument. This reflects a new drop in the centroid because the higher partials tend to fade before the lower ones, until the sound/note fades away, characterizing the offset. Figure 6 depicts the re gions (letters) and boundaries (numbers) of the ACT model for sustained (step-like excitation) sounds in part a) and an extension of the ACT model for percussive (impulse-like excitation) sounds in part b). In the figure, BN stands for background noise, A for attack, T for transition, S is sustain, D is decay, R is release

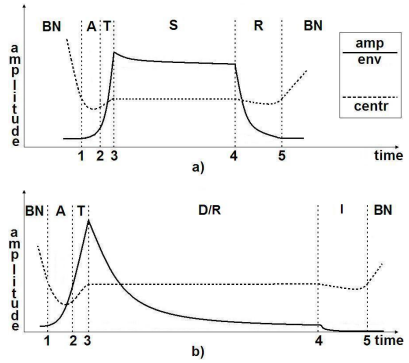

Figure 6: Temporal evolution of the spectral centroid and amplitude envelope for two classes of excitation modes. The figure shows the regions (letters) and their boundaries (numbers).

and I is interruption. The boundaries are the onset (1), end of attack (2), begin of transient (3), begin of release/interruption (4) and offset (5). Using this model, Hajda defines the attack as that part of the signal from onset during which the amplitude increases and the centroid decreases. Pre-attack noise is indicated by more or less uncorrelated fluctuations of both amplitude and centroid. According to the model, the attack ends when the centroid slope changes direction. A new segment, the attack/steady state transition, is defined as that segment immediately following the attack during which the amplitude continues to increase and the centroid increases overall. The sustain begins when the amplitude has achieved a local maximum; during this segment, the amplitude and centroid vary in a more or less monotonic fashion [14]. The release (or interruption) begins when both the amplitude and centroid decrease. Our approach uses the ACT model proposed by Hajda to automatically segment the temporal evolution of acoustic musical instrument sounds by detecting these five boundaries for as many types of sounds as possible, so we will test the model with sustained and percussive sounds, even though Hajda concluded that the ACT model does not seem to apply to most percussive sounds. Our technique uses the improved true amplitude envelope estimation that we developed instead of the RMS originally used [14].

6. AUTOMATIC SPECTRO-TEMPORAL SEGMENTATION

Here Figure 7 presents the regions automatically detected by our proposed method and compare our results with the baseline method by Peeters [10]. The importance of these figures is twofold. Firstly they should enable us to verify whether the ACT model works for the sustained (blown and bowed) and percussive (struck and plucked) sounds tested. Secondly, we will present the results of our automatic segmentation method and compare them to the baseline method. So we show the *fwr* waveform outlined by the TAE (solid line) and the centroid (dashed line) at the top and the corresponding spectrogram at the bottom. We included the spectrogram as an important visual aid in interpreting the segmentation results. That is, the spectrogram is intended to allow an intuitive visual inspection of the segments. The segmentation boundaries are shown as five solid vertical lines for our method and two dashed lines for the baseline method. The five boundaries correspond to those shown in Figure 6 while the two detected by the baseline method should represent what they de-

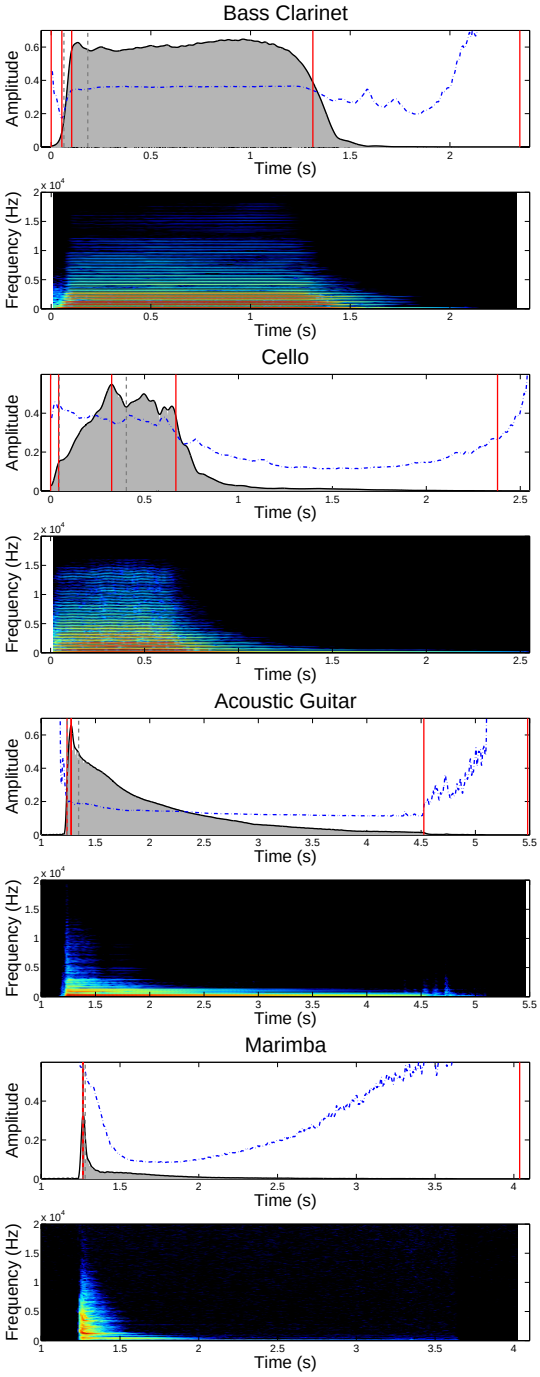

Figure 7: Full-wave rectified (*fwr*) waveform, TAE and centroid (top) and spectrogram (bottom) for sustained (blown and bowed) and percussive (struck and plucked) instruments.

fined as the attack (rise time), roughly corresponding to the onset and beginning of steady state in our model. We use an automatic onset detection method [26] to accurately define the first boundary, the onset (1). Then, the offset (5) is defined as the last point the TAE attains the same energy (amplitude squared) as the onset. The next step is the definition of the beginning of the sustain (3) and release (4) segments with a modified adaptive version of the effort method, explained bellow. Finally, we find the first local minimum between (1) and (3) and define it as the end of the attack (2). We should notice that we allow the same point to define the boundary of two distinct contiguous regions. This could happen remarkably during the transition between the end of the attack (2) and the beginning of sustain (3) and we interpret it as merely signifying that the transition is too short to be detected as a separate segment.

We employed adaptive efforts to independently measure the initial and final slopes because they exhibit typically different behaviors. This adaptive version is based on the original method of efforts proposed by Peeters [10] because it is more robust than the derivatives [9], [18], too sensitive to ripples. The adaptive effort is an indirect measure of the slope. We measure from the onset forward how many efforts summed are larger than the mean effort. Then we use this number as *M* in the original method. We do the same from the offset backwards. The results shown for the baseline method correspond to the original proposal [10].

We chose the instruments to try to display the result of the automatic segmentation for the four main classes considered, namely, blown pipe and bowed string (sustained), and plucked string and struck (percussive). These are challenging choices if we consider the original proposed application of the model and are meant to test its robustness. We will examine the results case by case. The clarinet (blown pipe) represents the best case scenario. It clearly fits the model perfectly and shows that the automatic segmentation accurately detects the boundaries because they are strongly present. Notably, the end of the attack (2), beginning of sustain (3), and begin of release (4) agree well with a visual inspection of the spectrogram. This seems to be the case for all blown instruments we tested, confirming the original model. We should notice that the baseline method detects both (2) and (3) with a considerable lag.

The cello shows that both the amplitude and the centroid might not behave like predicted for slow attacks typical of bowed strings. If we examine the spectrogram, the onset (1), release (4) and offset (5) were accurately detected. However, the end of the attack (2) is more difficult to determine, both visually and from the detection function. The beginning of the sustain is remarkably difficult to estimate for this case because neither the amplitude nor the centroid seem to confirm the visual inspection of the spectrogram. Bowed strings typically present a *crescendo* behavior reflected in the amplitude evolution that usually does not correspond to the vibration pattern measured by the centroid. Also, the centroid does not remain stable during the excitation mainly due to the *vibrato*, also typically present in bowed strings. It is evident, though, that the baseline method presents estimates that are far off. All bowed strings that we tested exhibited similar behavior.

The guitar shows that some plucked strings might exhibit a typical behavior that roughly fits the model. The steady vibrational pattern captured by the centroid is a typical example of how the decay cannot be characterized solely by the amplitude. However, since the attack is very fast, the boundaries between onset (1), end of attack (2) and beginning of decay (3) are blurred. This is a typical example of when (2) and (3) coincide. Also, it correctly detects the interruption (4), as previously explained. The baseline method also presents considerable detection lags in this case, overestimating the attack. Not all plucked strings tested, however, presented such clearly steady vibrational patterns that could be correctly identified by the model.

Finally, the marimba is a clear example of when the model assumptions break down. Even though the spectrogram would

lead us to visually guess the boundaries of some of the events, neither the centroid nor the amplitude envelope present a typical enough behavior to allow a correct detection of the regions according to the model. Notably, the vibration of the stick used to strike the key is clearly audible during the attack portion of the sound, constituting an important perceptual cue to the identification of struck percussions. When the vibrations of the instrument itself dominate, it is already too late to try and detect events. There are only two visible boundaries detected by our method. The onset (1) seems to be the only one correctly detected. End of attack (2) and beginning of decay (3) were bundled up. There is no beginning of interruption (4). In this case, the baseline method does not seem to find two separate points either, detecting the beginning too late. We would probably need to deconvolute the vibration of the stick from the key to be able to apply the model.

7. CONCLUSIONS AND FUTURE PERSPECTIVES

This work presents a method for automatically segmenting the temporal evolution of isolated acoustic musical instrument sounds using a model that takes spectro-temporal cues to correctly detect the boundaries of the regions. Even though the ACT model was originally proposed for sustained instrument sounds, we tested whether it can also be applied to some types of percussive sounds. An important contribution of this work is an improved amplitude envelope estimation technique (TAE) based on true envelope that optimally fits a curve that approximately matches the peaks of the waveform. TAE proved to be superior to the other methods tested. We found that the ACT model can be applied to plucked strings within certain boundaries. Finally, the automatic detection technique proposed performed better than the baseline method for most examples shown. We verified empirically that it outperforms the baseline method for all cases when the model fits well the type of instrument tested. More examples on http://recherche.ircam.fr/anasyn/caetano/seg.html.

Future perspectives of this work could include more robust detection techniques that potentially apply universally to all sounds, sustained or otherwise, or even separate automatic detection techniques devoted to each type of excitation.

8. ACKNOWLEDGMENTS

This work is supported by the Brazilian Governmental Research Agency CAPES (process 4082-05-2). The first author would like to thank Stephen McAdams for the inspiration.

9. REFERENCES

[1] Handel, S. “Timbre perception and auditory object identification.” In B.C.J. Moore (ed.), Hearing (pp. 425-461). New York: Academic Press, 1995.

[2] P. Iverson and C. L. Krumhansl, “Isolating the Dynamic Attributes of Musical Timbre,” J. Acoust. Soc. Am, 94(5), pp. 2595-2603, 1993.

[3] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., Krimphoff, J. “Perceptual Scaling of Synthesized Musical Timbres: Common Dimensions, Specifities and latent subject Classes”. Psychol. Res., 58, pp. 177-192, 1995.

[4] Peeters, G., McAdams, S., Herrera, P. Instrument Sound Description in the Context of MPEG-7. Proc. ICMC, 2000.

[5] N. Collins, “Investigating Computational Models of Perceptual Attack Time”. In Proc. ICMPC, 2006.

[6] J. W. Gordon, “The Perceptual Attack Time of Musical Tones,” J. Acoust. Soc. Am., 82(1), pp. 88-105, 1987.

[7] Schloss A. "On the Automatic Transcription of Percussive Music - From Acoustic Signal to High-Level Analysis". Ph.D. thesis, Stanford University, 1985.

[8] Bello , J.P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M.B. A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing. 13(5), Part 2, pages 1035-1047, September, 2005.

[9] K. Jensen. Envelope Model of Isolated Musical Sounds. Proc. DAFx, 1999.

[10] Peeters, G. “A large set of audio features for sound description (similarity and classification) in the CUIDADO project”. Project Report, 2004.

[11] Von Helmholtz, H. On the Sensations of Tone. London, Longman, 1885.

[12] Pinch, T and Trocco, F. Analog Days: The Invention and Impact of the Moog Synthesizer. Cambridge, MA: Harvard University Press, 2002.

[13] D. Luce and M. Clark. “Durations of Attack Transients of Nonpercussive Orchestral Instruments”. J. Audio Eng. Soc. 13 (3), pp. 194-199, 1965.

[14] J. Hajda “A New Model for Segmenting the Envelope of Musical Signals: The relative Salience of Steady State versus Attack, Revisited”, in Journal of the AES, Nov. 1996.

[15] A. Röbel, X. Rodet, “Efficient Spectral Envelope Estimation and its Application To Pitch Shifting And Envelope Preservation,” Proc. DAFx, 2005.

[16] A. Potamianos, P. Maragos. A Comparison of the Energy Operator and the Hilbert Transform Approach to Signal and Speech Demodulation. Signal Processing, 17 (1), pp. 95-120, 1994.

[17] Dodge, C. Jerse, T. A. Computer Music: Synthesis, composition and performance. Schirmer Books, Macmillan, New York, 1985. ISBN 0-02-873100-X.

[18] Skowronek, J., McKinney, M. “Features for Audio Classification: Percussiveness of Sounds” in Intelligent Algorithms in Ambient and Biomedical Computing. Springer Netherlands, 2006.

[19] M. Athineos, D. P. W. Ellis, “Frequency-Domain Linear Prediction for Temporal Features.” Proc. IEEE ASRU Workshop, 2003.

[20] G.Tzanetakis, P.Cook, Sound analysis using MPEG compressed audio, In Proc. IEEE ICASSP, pp.761-764, Vol.2, Istanbul, Turkey, 2000.

[21] Xu, C., Zhu, Y., and Tian, Q., Automatic music summarization based on temporal, spectral and cepstral features, In Proc. IEEE ICME, pp. 117-120, Lausanne, Switzerland, 2002.

[22] Xi Shao, Changsheng Xu, Ye Wang, Mohan S Kankanhalli Automatic Music Summarization in Compressed Domain Proc. ICASSP 2004.

[23] Makhoul, J. “Linear prediction: A tutorial review” Proc. IEEE, vol. 63, pp. 561-580, Apr. 1975.

[24] Risset, J.C., & Wessel, D.L. (1982). Exploration of timbre by analysis and synthesis, *The Psychology of Music*, D. Deustch (Ed.). Orlando, FL: Academic Press, 26-58.

[25] Risset, J. C. Computer Study of Trumpet Tones. Murray Hill, N.J.: Bell Telephone Laboratories, 1966.

[26] A. Röbel, A New Approach to Transient Processing in the Phase Vocoder. Proc. DAFx, 2003.