Proceedings of Meetings on Acoustics

Volume 15, 2012 http://acousticalsociety.org/

163rd Meeting Acoustical Society of America/ACOUSTCS 2012 HONG KONG

Hong Kong
13 - 18 May 2012

Session 4aMU: Musical Acoustics

4aMU1. Real-time segmentation of the temporal evolution of musical sounds

John Glover*, Victor Lazzarini and Joseph Timoney

***Corresponding author's address: The Sound and Digital Music Research Group, National University of Ireland, Maynooth, Music Technology Laboratory 1, Maynooth, n/a, Co. Kildare, Ireland, john.c.glover@nuim.ie**

Since the studies of Helmholtz, it has been known that the temporal evolution of musical sounds plays an important role in our perception of timbre. The accurate temporal segmentation of musical sounds into regions with distinct characteristics is therefore of interest to researchers in the field of timbre perception as well as to those working with different forms of sound modelling and manipulation. Following recent work by Hajda (1996), Peeters (2004) and Caetano et al (2010), this paper presents a new method for the automatic segmentation of the temporal evolution of isolated musical sounds in real-time. We define attack, sustain and release segments using cues from a combination of the amplitude envelope, the spectro- temporal evolution and a measurement of the stability of the sound that is derived from the onset detection function. We conclude with an evaluation of the method.

Published by the Acoustical Society of America through the American Institute of Physics

Received 7 Jun 2012; published 3 Aug 2012

1. INTRODUCTION

The segmentation of musical instrument sounds into contiguous regions with distinct characteristics has become an important process in studies of timbre perception [1] and sound modelling and manipulation [2]. Since the time of Helmholtz, it has been known that the temporal evolution of musical sounds plays an important role in our perception of timbre. Helmholtz described musical sounds as being a waveform shaped by an amplitude envelope consisting of attack, steady state and decay segments [3]. Here the attack is the time from the onset until the amplitude reaches its peak value, the steady state is the segment during which the amplitude is approximately constant, and the decay is the region where the amplitude decreases again. A number of automatic segmentation techniques have been developed based on this model, taking only the temporal evolution of the signal amplitude into account when calculating region boundaries [4, 5]. However, more recent research has shown that in order to better understand the temporal evolution of sounds, it is necessary to also consider the way in which the audio spectrum changes over time [1]. In [6] Hajda proposed a model for the segmentation of isolated, continuant musical tones based on the relationship between the temporal evolution of the amplitude envelope and the spectral centroid, called the amplitude/centroid trajectory (ACT) model. Caetano et al. devised an automatic segmentation technique based on the ACT model and showed that it outperformed a segmentation method based solely on the temporal evolution of the amplitude envelope [7]. While this method works well in an off-line situation, it cannot be used to improve real-time systems. We are particularly interested in real-time musical performance tools based on sound synthesis by analysis, where the choice of processing algorithm will often depend on the characteristics of the sound source. Spectral processing tools such as the Phase Vocoder [8] are well established means of time-stretching and pitch-shifting harmonic musical notes, but they have well documented weaknesses in dealing with noisy or transient signals [9]. For real-time applications of tools such as the Phase Vocoder, it may not be possible to depend on any prior knowledge of the signal to select the processing algorithm, so being able to accurately identify note regions with specific characteristics on-the-fly is crucial in order to minimize synthesis artifacts.

In this paper, we present a new technique for the real-time automatic temporal segmentation of musical sounds. We define attack, sustain and release segments using cues from a combination of the amplitude envelope, the spectral centroid, and a measurement of the stability of the sound that is derived from an onset detection function. In Section 2 we describe some existing approaches to automatic segmentation. Our new method is given in Section 3. We provide an evaluation of our method in Section 4, followed by conclusions in Section 5.

2. AUTOMATIC SEGMENTATION

Automatic segmentation consists of the identification of boundaries between contiguous regions in a musical note. Typical boundaries are at note onsets, the end of the attack or the start of the sustain, the end of the sustain or the start of the release and at the end of the note (offset). Regions and boundaries can vary however, firstly depending on the model used by the segmentation technique and secondly based on the nature of the sound being analyzed, as not all instrumental sounds are composed of the same temporal events. In [4] Peeters notes that the well-known ADSR envelope does not apply to most natural sounds, as depending on the nature of the sound, one or more of the segments is often missing. Therefore, he proposes segmenting musical sounds into two regions named *attack* and *rest*. This only requires the detection of two region boundaries; the start and end of the attack region. Two techniques are described for detecting these boundaries. The first of these is just to apply a fixed threshold to the amplitude envelope, the start of attack being when the envelope reaches 20% of the peak value and the end of attack occurring when it reaches 90% of the peak value. The second technique is called the *weakest effort* method, and is based on an indirect measurement of the changes in the slope of the amplitude envelope.

Although the amplitude envelope often provides a good approximation of the temporal evolution of the internal structure of a musical sound, it simply does not provide enough information to allow for accurate, robust and meaningful segmentation of the signal into regions with distinct characteristics. In particular the attack region, which has often become synonymous with the amplitude rise time [10], is not well delineated by the amplitude envelope. The attack is a transient part of the signal that lasts from the onset until a relatively stable periodicity is established, and as a result the steady state is generally achieved before the end of the initial amplitude rise time [1]. During the steady state, the amplitude envelope can often show considerable variation, particularly in the presence of tremolo and/or vibrato. This makes it difficult to detect the boundary between the steady state and the release using just the amplitude envelope, especially if operating under the constraints of a real-time system. The ACT model, which we describe in Section 2.1, has addressed many of these issues.

2.1. Automatic segmentation using the amplitude/centroid trajectory model

Hajda proposed a new model for the partitioning of isolated non-percussive musical sounds [6], based on observations by Beauchamp that for certain signals the root mean square (RMS) amplitude and spectral centroid have a monotonic relationship during the steady state region [11]. An example of this relationship is shown for a clarinet sample in Figure 1. The spectral centroid is given by Equation 1, where f is frequency (in Hz) and a is linear amplitude of frequency band b up to m bands which are computed by Fast Fourier Transform. The Fourier Transform is performed on Bartlett windowed analysis frames that are 64 samples in duration. This results in 32 evenly spaced frequency bands (up to 11025 Hz), each with a bandwidth of about 345 Hz.

$$centroid(t) = \frac{\sum_{b=1}^{m} f_b(t) \times a_b(t)}{\sum_{b=1}^{m} a_b(t)} \tag{1}$$

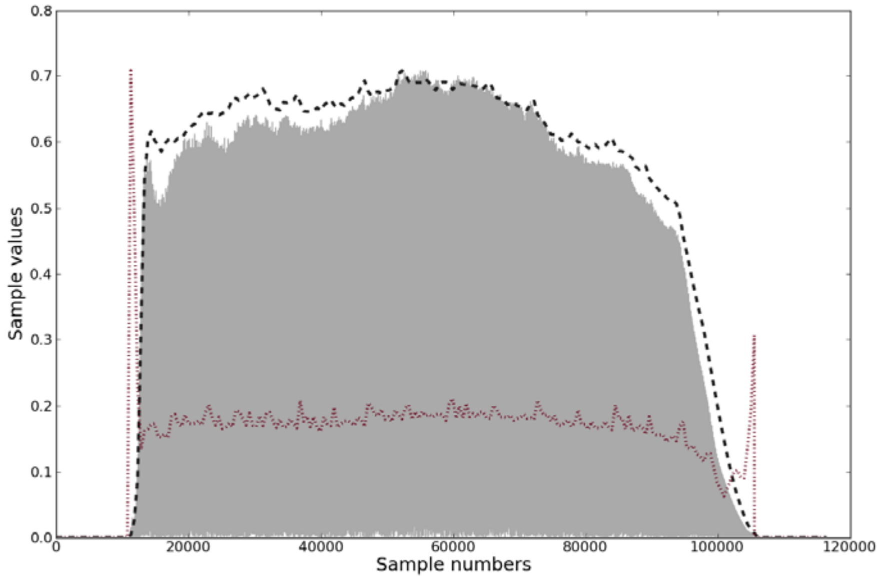

FIGURE 1. The relationship between the RMS amplitude envelope and spectral centroid for a clarinet sample. The full-wave-rectified version of the audio sample is given, with the RMS amplitude envelope shown by the black dashes, and the spectral centroid by the red dots. The RMS amplitude envelope and the spectral centroid have both been normalised and scaled by the maximum signal value.

Hajda's model, called the amplitude/centroid trajectory (ACT), identifies and detects the boundaries for four contiguous regions in a musical tone:

Attack: the portion of the signal in which the RMS amplitude is rising and the spectral centroid is falling after an initial maximum. The attack ends when the centroid slope changes direction (centroid reaches a local minimum).

Attack/steady state transition: the region from the end of the attack to the first local RMS amplitude maximum.

Steady state: the segment in which the amplitude and spectral centroid both vary around mean values.

Decay: the section during which the amplitude and spectral centroid both rapidly decrease. At the end of the decay (near the note offset), the centroid value can rise again however as the signal amplitude can become so low that denominator in Equation 1 will approach 0. This can be seen in Figure 1 (starting at approximately sample number 100200).

Hajda initially applied the ACT model only to non-percussive sounds. However, Caetano et al. introduced an automatic segmentation technique based on the ACT model [7], and proposed that it could be applied to a large variety of acoustic instrument tones. It uses cues taken from a combination of the amplitude envelope and the spectral centroid, where the amplitude envelope is calculated using a technique called the true amplitude envelope (TAE) [12]. The TAE is a time domain implementation of the true envelope [13], which is a method for estimating a spectral envelope by iteratively calculating the filtered cepstrum, then modifying it so that the original spectral peaks are maintained while the cepstral filter is used to fill the valleys between the peaks. In the TAE this algorithm is applied to the time domain signal instead of the Fourier spectrum, so that the resulting envelope accurately matches the time domain amplitude peaks.

For each musical tone the onset, end of attack, start of sustain, start of release and offset boundaries are detected as follows:

Onset: start of the note, found by using the automatic onset detection method described in [14]. This technique basically involves looking for signal regions in which the center of gravity of the instantaneous energy of the windowed signal is above a given threshold. Or in other words, if most of the energy in a spectral frame is located towards the leading edge of the analysis window, then the frame is likely to contain a note onset.

End of attack: position of the first local minima in the spectral centroid that is between boundaries 1 and 3.

Start of sustain: boundary detected using a modified version of Peeters' weakest effort method.

Start of release: also detected using a version of the weakest effort method, but starting at the offset and working backwards.

Offset: the last point that the TAE attains the same energy (amplitude squared) as the onset.

Notably, they allow the same point to define the boundary of two distinct contiguous regions. This signifies that the region is too short to be detected as a separate segment and makes the model more robust in dealing with different types of sounds.

Caetano et al. compare the performance of their automatic segmentation technique to that of the one described by Peeters [4]. They do this by visual inspection of plots of the waveform, spectrogram and detected boundaries produced by both methods, showing 16 analyzed samples consisting of isolated tones from western orchestral instruments (plus the acoustic guitar). They find that their model outperformed the Peeters method in all cases, although for one sample (a marimba recording) the amplitude envelope and spectral centroid do not behave in the manner that is assumed by the model, and so neither method gives good results. However, this provides strong evidence that the ACT model assumptions can be applied to a wide variety of sounds, and shows that using a combination of the amplitude envelope spectral centroid can lead to more accurate note segmentation than methods based on the amplitude envelope alone.

The automatic segmentation technique proposed by Caetano et al. cannot be used to improve the performance of real-time synthesis by analysis systems however, as the method for detecting the start of sustain and start of release boundaries requires knowledge of future signal values. Also, although the spectral centroid has been shown to be a useful indirect indicator as to the extent of the attack region, in order to help reduce synthesis artifacts when using tools such as the Phase Vocoder it would be preferable to have a more accurate measure of the attack transient, by locating the signal regions in which the spectral components are changing rapidly and often unpredictably. We address both of these issues in Section 3.

3. A REAL-TIME METHOD FOR THE AUTOMATIC TEMPORAL SEGMENTATION OF MUSICAL SOUNDS

In this section we propose a new method for the real-time automatic segmentation of the temporal evolution of musical sounds, using cues from a combination of the RMS amplitude envelope, the spectral centroid and an onset detection function (the latter is described in Section 3.1). In our segmentation model, boundaries are defined for the onset, start of sustain, start of release and offset as follows:

Onset: start of the note, detected using the peak amplitude difference method [15].

Start of sustain (end of attack): boundary occurs as soon as the attack transient has finished. This calculation is described in detail in Section 3.1.

Start of release (end of sustain): occurs when the following conditions are met:

(1) The RMS amplitude envelope is less than 80% of the largest amplitude value seen between the onset and the current frame.

(2) The RMS amplitude envelope is decreasing for 5 consecutive frames.

(3) The current value of the spectral centroid is below the cumulative moving average of the values of the centroid from the onset to the current frame.

This boundary also occurs if the RMS amplitude value drops to less than 33% of the peak value. The RMS amplitude here is subject to a 3 point moving average filter, and the spectral centroid is given by Equation 1.

Offset: the point at which the RMS amplitude value is 60 db below the peak amplitude value.

Similarly to the Caetano et al. method [7], we allow multiple boundaries to occur at the same time. An example of the boundaries detected by our method is given in Figure 3, with boundary positions shown by vertical blue dashes. We use a frame size of 512 samples, resulting in a latency of 11.6 ms when operating at a sampling rate of 44.1 kHz. The maximum delay in detecting a boundary is 5 frames (or 58 ms).

3.1. Identifying the attack transient from an onset detection function

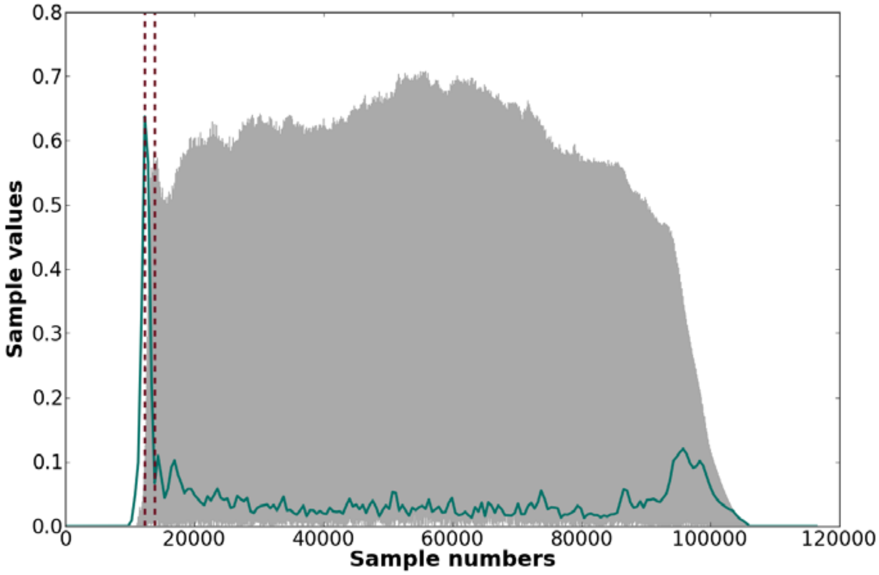

FIGURE 2. Onset detection function (solid green line) for the clarinet sample, and detected transient region (between red vertical dashes).

As onset locations are typically defined as being the start of a transient, the problem of finding their position is linked to the problem of detecting transient regions in the signal. Another way to phrase this is to say that onset detection is the process of identifying which parts of a signal are relatively unpredictable. The majority of the onset detection algorithms described in the literature involve an initial data reduction step, transforming the audio signal into an *onset detection function* (ODF), which is a representation of the audio signal at a much lower sampling rate. The ODF usually consists of one value for every frame of audio and should give a good indication as to the measure of the unpredictability of that frame. Higher values correspond to greater unpredictability. The onset detection function used in our model is the peak amplitude difference method, one of the best-performing methods discussed in [15]. It is based on the premise that during the steady state of a musical note, a quasi-harmonic signal can be well modelled as a sum of sinusoidal partials with slowly evolving amplitudes, frequencies and phases. Therefore, the absolute values of the frame-to-frame differences in the sinusoidal peak amplitudes and frequencies should be quite low. In comparison, transient regions at note onset locations should show considerably more frame-by-frame variation in both peak frequency and amplitude values, so an ODF can be created by measuring these frame-by-frame amplitude and frequency variations. As this effectively measures errors in the partial tracking stage of sinusoidal modelling [17], it can also be used to measure the stability of the detected sinusoidal partials in the audio signal. Peaks in the ODF should therefore occur at regions where the spectral components in the signal are most unstable or are changing unpredictably. It should be noted that this does not only apply to our ODF, but indeed any ODF that measures the variability of spectral components in the audio signal.

We define the attack transient as being the region from the onset until the next local minimum in the ODF. Additionally, we also signal the end of the attack segment if the RMS amplitude envelope reaches a local maxima. This technique is similar to the transient detection method proposed in [9], where the authors detect transient regions based on peaks in the energy of the noise signal resulting from the identification and removal of the deterministic signal component. However, as we do not do a separation of the deterministic component from the stochastic, our method should require considerably less computation. In addition, we do not low-pass filter the resulting ODF, as doing so widens the ODF peak (and in turn, the detected transient region), without presenting an obvious way to compensate for this deviation. An example of the ODF and corresponding transient region can be seen in Figure 2.

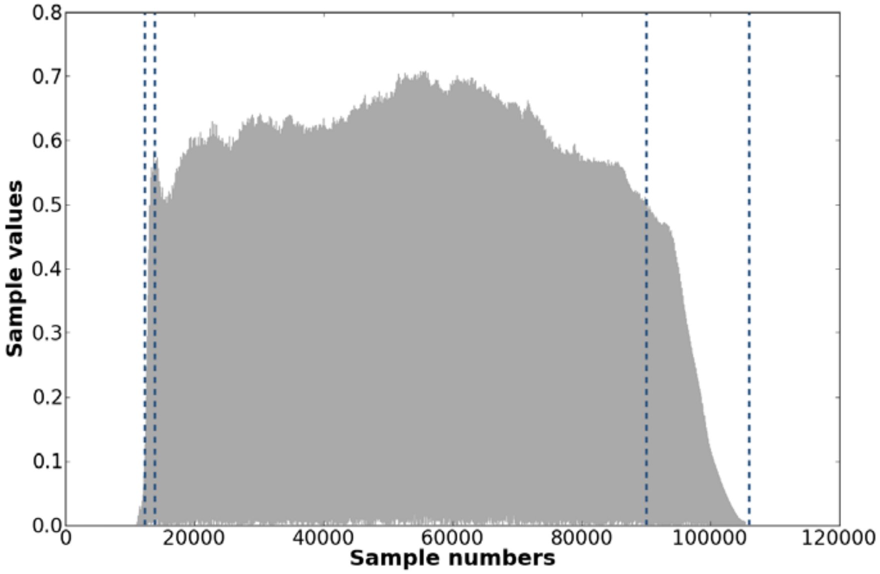

FIGURE 3. Boundaries detected by our proposed real-time segmentation method. The boundaries (from left to right) are the onset, start of sustain (end of attack), start of release (end of sustain) and offset.

4. RESULTS

To evaluate the performance of our proposed segmentation model, we compared the locations of the detected boundaries with those found by our implementation of the method given by Caetano et al. [7]. We took a selection of 36 samples of isolated musical sounds from the Modal database [15], which is a freely available database of samples with creative commons licensing allowing for free reuse and redistribution. More information about the Modal database can be found at `http://www.johnglover.net`. The samples are quite diverse in nature, covering percussive and non-percussive sounds from a mixture of western orchestral instruments, contemporary western instruments and vocal samples. Initially created to evaluate the performance of real-time onset detection algorithms, Modal includes hand-annotated onset locations for each sample. For this work, three additional annotations were added for each sample: start of sustain, start of release and note offset. The annotations were all made by one person, which will inevitably lead to some degree of inaccuracy and inconsistency as is shown in [16], however they should still give some indication as to the performance of the automatic segmentation methods. In addition to the hand-annotated boundaries, we also developed an automatic technique for identifying regions in the audio signal with the highest level of sinusoidal partial instability. This was done by first performing sinusoidal analysis on each sample using the Spectral Modelling Synthesis method [17], then calculating a detection function from the sum of the frame by frame variations in log frequency (scaled by log amplitude) for each partial. Areas with unstable partials were then defined as the area around peaks in this detection function. The automatically detected segmentation boundaries were compared to each of the hand-annotated boundaries plus the additional partial instability region measurement.

Table 1 gives the average difference in milliseconds between the automatically detected boundary and reference boundary for our method and the Caetano et al. method. The onset detection results are identical, as we used the same onset detection algorithm for both methods (the peak amplitude difference technique). The Caetano et al. method sustain boundary is slightly closer to the hand-annotated sustain locations on average (by 2.2 ms), but our sustain section is considerably closer to the end of the region with the highest level of partial instability. Our method also fares better in detecting start of release and note offset locations in comparison with the Caetano et al. method. A large part of the error in the Caetano et al. offset detection can be attributed to the fact that they define this boundary based on the energy the signal has at the onset location, and as our onset detector is a real-time method there is a slight latency before it responds, by which stage the signal energy has already started to increase.

TABLE 1. Average deviation from boundaries in reference samples for our proposed method and for the Caetano et al. method.

Boundary	Proposed Method Avg. Deviation (ms)	Caetano et al. Method Avg. Deviation (ms)
Onset	16.6	16.6
Start of sustain	67.1	64.9
End of unstable partials	40.5	80.0
Start of release	541.6	900.7
Offset	329.5	1597.7

TABLE 2. Model accuracy for our proposed method and for the Caetano et al. method.

Boundary	Proposed Method Accuracy (%)	Caetano et al. Method Accuracy (%)
Onset	97.2	97.2
Start of sustain	83.3	77.8
End of unstable partials	91.7	69.4
Start of release	30.6	38.9
Offset	58.3	25.0

When evaluating onset detection algorithms, an onset is commonly regarded as being correctly detected if it falls within 50 ms of the reference onset location in order to allow for human error when creating the set of reference values [10, 16]. As the note segmentation boundaries are often a lot more difficult to annotate accurately (the start of the sustain section in particular is not easy to define), we have allowed a more lenient detection window of 100 ms. Table 2 gives the percentage of automatically detected boundaries that fall within 100 ms of the reference values for both segmentation methods. Here, our proposed method is slightly more accurate in detecting the sustain boundary, but our sustain section is again considerably closer to the end of the unstable partial region. The Caetano et al. method is more accurate in detecting the release, with our method performing better at note offset detection.

The results show that both methods perform reasonably well at detecting the start of the sustain region, although our start of the sustain region is significantly closer to the end of the region with high partial instability. Neither method performs particularly well in detecting the release and offset with high accuracy, although on average our proposed model behaves more robustly. Our model is also suitable for real-time use unlike the Caetano et al. method.

5. CONCLUSIONS

This paper proposed a new model for the real-time segmentation of the temporal evolution of musical sounds, using cues from the amplitude envelope, spectral centroid and an onset detection function that is based on measuring errors in sinusoidal partial tracking. We evaluated our method by comparing it with the technique proposed by Caetano et al. and found that in the average case it generally performs better and is more robust. Our method can run in real-time and with considerably lower computation requirements as it does not calculate the computationally costly true amplitude envelope. Neither method was particularly accurate in detecting the release and offset boundaries, so future work could include some research in this area. We will also work on integrating the segmentation system with a real-time performance tool based on sinusoidal synthesis by analysis. The code for our segmentation method, our reference samples and all of the code needed to reproduce our results can be found online at
`http://www.johnglover.net`.

ACKNOWLEDGMENTS

The authors are grateful for the financial support that was provided by the Postgraduate Travel Fund at the National University of Ireland, Maynooth.

REFERENCES

1. J. M. Hajda, "The Effect of Dynamic Acoustical Features on Musical Timbre", in Analysis, Synthesis, and Perception of Musical Sounds, J. W. Beauchamp, ed. (Springer, New York), pp. 250-271, 2007.
2. M. Caetano and X. Rodet, "Automatic Timbral Morphing of Musical Instrument Sounds by High-Level Descriptors", Proc. 2010 Int. Computer Music Conf. (ICMC 2010), New York, 2010.
3. H. V. Helmholtz, "On the Sensations of Tone as a Physiological Basis for the Theory of Music", Dover, New York, 1877.
4. G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification)", The CUIDADO Project, Project Report, 2004. `http://recherche.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf` (last accessed 18-07-2012).
5. K. Jensen, "Envelope Model of Isolated Musical Sounds", Proc. 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), Trondheim, Norway, 1999.
6. J. M. Hajda, "A New Model for Segmenting the Envelope of Musical Signals: The Relative Salience of Steady State Versus Attack, Revisited", Audio Eng. Soc. Paper No. 4391, 1996.
7. M. Caetano, J. J. Burred, and X. Rodet, "Automatic Segmentation of the Temporal Evolution of Isolated Acoustic Musical Instrument Sounds Using Spectro-Temporal Cues", Proc. 13th Int. Conf. on Digital Audio Effects (DAFx-10), Graz, Austria, 2010.
8. M. Dolson, "The Phase Vocoder: A Tutorial", Computer Music Journal, 10 (4), pp. 14-27 (1986).
9. C. Duxbury, M. Davies and M. Sandler, "Improved Time-Scaling of Musical Audio Using Phase Locking at Transients", 112th Audio Eng. Soc. Convention, Paper No. 5530, Munich, Germany, 2002.
10. J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. Sandler, "A Tutorial on Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing, 13, pp. 1035-1047 (2005).
11. J. Beauchamp, "Synthesis by Spectral Amplitude and 'Brightness' Matching of Analyzed Musical Instrument Tones" Journal of the Audio Eng. Soc., 30, pp. 396-406 (1982).
12. M. Caetano and X. Rodet, "Improved Estimation of the Amplitude Envelope of Time-Domain Signals Using True Envelope Cepstral Smoothing", Proc. 2011 Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2011), pp. 4244-4247, Prague, Czech Republic, 2011.
13. A. Röbel and X. Rodet, "Efficient Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation", Proc. 8th Int. Conf. on Digital Audio Effects (DAFx-05), Madrid, Spain, 2005.
14. A. Röbel, "A New Approach to Transient Processing in the Phase Vocoder", Proc. 6th Int. Conf. on Digital Audio Effects (DAFx-03), London, UK, September, 2003.
15. J. Glover, V. Lazzarini and J. Timoney, "Real-Time Detection of Musical Onsets with Linear Prediction and Sinusoidal Modeling", EURASIP Journal on Advances in Signal Processing, 2011 (1), pp. 68 (2011).
16. P. Leveau, L. Daudet and G. Richard, "Methodology and Tools for the Evaluation of Automatic Onset Detection Algorithms in Music", Proc. 5th Int. Conf. on Music Information Retrieval (ISMIR 2004), pp. 72-75, Barcelona, Spain, 2004.
17. X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition", Computer Music Journal, 14 (4), pp. 12-24 (1990).