METHODS AND APPLICATIONS FOR SEGMENTATION AND DESCRIPTION OF MUSICAL TRANSIENTS

Sarah Smith
University of Rochester
Department of Electrical and Computer Engineering
ECE 492 - Literature Review Draft

ABSTRACT

Separating an audio signal into regions corresponding to transients and steady state is an important step in many applications. Perceptual studies have found that humans often perceive and process transients separately from steady state tones. This distinction also has a basis in most physical systems, such as instruments, which have a different response to sustained as opposed to impulsive excitations. Furthermore, transient regions often display very different signal properties than steady state sections. Therefore, many audio encoders use this type of segmentation as a preprocessing step and adjust the encoding methods accordingly. However, despite the importance of transients to many musical processes, the exact definition of a transient can be hard to define, resulting in a lack of ground truth estimate that can be used to compare algorithms. After discussing a few different definitions of musical transients and their applications, this paper provides an overview of note segmentation algorithms that aim to isolate the transient portions of a musical note.

1. INTRODUCTION

When analyzing audio, it is common to devote the most attention to the stable portions of the signal. For many applications, such as pitch detection or instrument identification, the steady state response of the instruments are generally sufficient for characterization. Processing these regions generally gives more consistent results than attempting to address the more varied characteristics of musical onsets. However, it is widely known that humans are very sensitive to transients in audio, with the length and type of transient often being used to identify different instrument timbres [10] and localize sound. Furthermore, performing musicians often devote careful attention to the attack and release portions of notes. As such, it is useful to consider how methods for identifying and characterizing musical transients could be useful in audio analysis applications.

1.1 Definition of a musical transient

However, as with many musical characteristics, there does not exist a single definition of 'transient' that can be used across applications. One could use a dictionary as a starting point, however, it does not contain enough description to be a useful working definition. The Oxford English Dictionary defines the adjective 'transient' as meaning: "Passing by or away with time; not durable or permanent; temporary, transitory" [17]. Although this definition is quite general, it does highlight some of the key distinctions that are present in audio. In particular, it contrasts the notion of transientness with that of a more permanent steady state. Additionally, the definition suggests that transients should constitute the minority of a signal, occurring at points of change and then passing away. Given the vague nature of this definition, different applications often choose contrasting and even conflicting metrics to determine the presence of a transient. As such it is often difficult to establish an agreed upon metric for determining a ground truth measurement against which transient detectors can be compared.

Furthermore, there is often an implicit distinction in how transients are defined depending on whether the application seeks to isolate or remove the transient sections of audio. In particular, applications that seek to isolate only the steady state portion of a tone for further analysis can often tolerate a more relaxed definition that might risk extending the transient label to include early portions of the steady state. [10] However, in cases where the transient portion of the signal is being explicitly analyzed, a more specific definition is often used, that may relate to expectations based on a source instrument model or similar assumptions [13].

1.2 Relation to Onset Detection

The notion of an auditory transient is closely related to that of musical onsets, due to the fact that the most perceptually relevant transients are often associated with the start of individual notes. During this initial stage, the instrument, or other sound source, rapidly changes from a static state to a periodically vibrating one. Also, unlike musical offsets, which tend to decay in a somewhat predictable manner, the onsets of musical notes are far more varied and often carry more musical importance. As such, many of the commonly used onset detection algorithms have applications in transient detection as well. However, there is an important distinction between these two tasks. Onset detection is generally associated with identifying a single point in time that corresponds to the perceived start of the note. In contrast, transient detectors seek to isolate a region of audio for encoding or analysis. If occurring at the beginning of a musical note, this region will often include the perceived onset as well as some amount of signal before and after it. In fact, many onset detection algorithms use a feature based detection function and select onset times based on the peak locations of this detection function. [1] Many algorithms of this type could therefore be adapted to the task of transient detection simply by choosing an appropriate threshold compared to the local maximum and identifying the region where the detection function rises above this level. With some regard to these similarities, this paper will focus primarily on methods for identifying transient segments within a recorded sound. For a comprehensive overview of onset detection methods, including many not discussed here, the reader should consult the earlier paper by Juan Bello et al. [1].

2. APPLICATIONS OF TRANSIENT DETECTION

Given the perceptual distinction between transients and steady state signals, it is often useful or necessary to encode these regions in different ways. For instance, applications that attempt to effectively parametrize audio signals for compression or encoding often use very different models for transient regions than they do for the sustained portions of notes. Similarly, when altering the pitch or tempo of a musical recording, it is often not necessary to change the length or pitch of the transient region. In both of these cases the primary emphasis is placed on labelling sections of audio as either 'transient' or 'sustain', with little need to further describe or characterize the transient region. Alternatively, many performance practice applications are interested in explicitly characterizing the initial portion of a note in relation to the instrument being played. In these applications, the aim is to understand how a musician affects the sound of their instrument in subtle ways. Although all of these methods rely on some form of transient detection, they all seek to extract different information from the signal and often modify the transient detection criteria accordingly.

2.1 Perceptual Audio Encoding

The purpose of an audio encoder is to reduce the bit rate of an audio file as much as possible without introducing undesirable amounts of audible distortion. This data compression is generally accomplished by exploiting psychoacoustic masking in the frequency domain in order to dynamically allocate bits within a frame. Within this context, frame lengths on the order of 10 milliseconds are often used in order to achieve the necessary frequency resolution. However, the sudden changes that are present during transient regions often require the use of shorter frames to get the necessary time resolution [2]. Dynamically adjusting this frame size in the course of the audio file allows for the optimal time frequency resolution to be used at each point in the sound, increasing the overall perceptual quality for a given bit rate. In these encoders, a transient detector is often used in order to determine how long of a block length to use when encoding a given section of audio. Unlike the segmentations used in time morphing and performance analysis, these detectors do not require each note to consist of a single attack phase followed by a sinusoidal steady state, but allow for the signal to change back and forth between these states relatively quickly [4]. This flexibility influences the choice of transient detector, often favoring spectral methods such as a high frequency energy detector [9].

2.2 Time and Frequency Morphing

In applications where the resynthesized sound is morphed in time or frequency, it is often desirable to first remove the transient portions before performing the modification. [7]. Although a pitch shifting or time stretching can be accomplished by applying the same process, usually a phase vocoder, to the entire sound, it is often more perceptually relevant to only alter the steady state portions of each note. [7]. Since non percussive instruments are capable os sustaining a note for an arbitrary length of time after an initial onset, it is logical to assume that the onset is not heavily influenced by the eventual note duration. Similarly, since many onset transients have a non harmonic frequency spectrum, they are generally less affected by changes in pitch than the harmonic portions of the note [1]. As such, many time morphing algorithms use a transient detector to identify regions for note onsets that are not time stretched, but simply reinserted in their original form at the appropriate points after the steady state signals have been morphed [7]

2.3 Performance Analysis

Musical transients can serve an important role in understanding how musicians perform different styles of music. Even when two performers are playing the same notes, rhythms, and dynamics, the resulting sounds can differ greatly. In attempting to identify these differences the transients are often very important [11] [16]. As such, many performance analysis applications are interested in segmenting notes into regions of attack, sustain, and release. As such, the detection algorithms are usually more complex than some other algorithms and are designed to provide more information about the time evolution of the note beyond identifying the segmentation boundaries. While some applications focus almost entirely on the amplitude envelope of the signal to assign boundaries, it is more common to combine the amplitude with the trajectory of one or more spectral features [11] [3].

3. METHODS FOR TRANSIENT DETECTION

Unlike the steady state portion of musical notes, onsets and offsets are often characterized by a broadband, generally non harmonic, frequency spectrum. Additionally, transients associated with the onsets or offsets of notes often correspond to changes in the overall energy of the signal. When used separately or together, these characteristics can be used to accurately isolate regions of transientness in a signal. While many segmentation algorithms use both the time and frequency domain information from the signal, there are also cases where only one of these domains is necessary. As such, the existing methods for transient identification can be broadly grouped according to whether they use a time domain, frequency domain, or combined time-frequency representation of the signal for analysis.

3.1 Time Domain Approaches

Perhaps the simplest way of isolating transients is to define a constant time from the start of a note that is considered to be transient. This value is usually chosen in the range of 60-100 ms and was used in many early perceptual studies [5]. In addition to the ease of implementation, this method can be useful when there is a need to standardize the length of transients across different instruments. In many of these studies, the researchers were attempting to identify the relevance of the attack phase in perception by separating the note into two segments that could be presented independently of each other. In this case it is necessary to normalize the length of the tones presented as attacks. Additionally this method can be useful when the intention is to sample a representative section of steady state or transient. It is straight forward to implement and can be biased towards eliminating any transient by choosing a longer value for the onset, or guaranteeing that no portion of the steady state is labelled as a transient by choosing a shorter value, depending on the needs of the experiment.

For more refined applications, however, the assumption of a constant attack time is often quite inaccurate. In fact, variations in attack time have been shown to greatly influence our perception of musical sounds [10] [5]. It is natural to expect the transient length to change depending on the instrument, pitch, and overall volume of the note as well as other factors. As such, many methods define the duration of a transient in relation to the change in energy of a signal.Under this definition, transients are considered to be regions with a rapid increase or decrease in signal energy. To measure the signal energy, the root mean square amplitude is generally used.

$$Amp = \frac{1}{N}\sum_{n=0}^{N-1}\sqrt{|x(n)|^2} \tag{1}$$

The frame length, N, can vary depending on the application, but common values are on the order of 20 ms. In general however, the resulting trajectory can be quite noisy. As such, it is generally necessary to apply a low pass filter to the calculated RMS values before continuing with further analysis. Once a suitable amplitude trajectory has been calculated, the question remains as to the best definition of the attack. One possibility is to define the end of the attack as the point of maximum amplitude. This choice has its basis in model used in many synthesizers, which partition the amplitude of each note into regions of attack, decay, sustain, and release. This encoding allows for the entire shape of the note to be encoded in 4 parameters, the attack rate, decay rate, sustain level, and release rate. These values are then combined with the overall length and loudness of a note to generate different timbres.

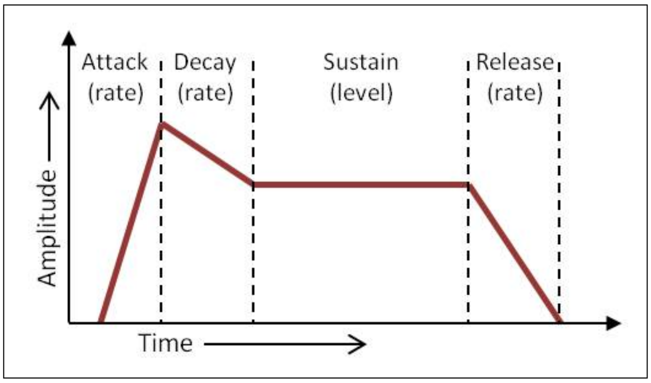

Figure 1. ADSR envelope model

Due to its historical relevance, this type of segmentation remains useful in synthesis applications, whose eventual aim is to imitate the recorded sound with a synthesizer. However, audio analysis applications tend to avoid this method since it is so heavily dependant on the location of a single peak. In recorded signals, it is likely that there may be an anomalous peak in amplitude later in the sound, or that the signal will exhibit very little decay. In either of these cases, an detector that uses only information about the maximum amplitude could detect the end of the transient far later than it actually occurs. As such, many amplitude based detectors define the attack phase as ending when the amplitude reaches a specified fraction of the maximum. The most common value for this threshold is 50 percent of the final amplitude, corresponding to a level of -3 dB, since it almost always occurs during the initial attack of the sound and includes most of the transient [10]. While using a larger value for this threshold would lengthen the region of the identified transient to include more of the signal energy rise, it also increases the chance of an erroneous detection.

Time domain and amplitude based methods are popular for many applications for their ease of implementation. They are robust to the most common sources of signal variation while making very few assumptions about the nature of the sound. In addition, they can be implemented in an on-line system with very low latency, unlike some frequency domain methods which require additional processing. However, these methods must estimate, or be given, a start time for the note, which is subject to many of the same sources of error as detecting the end of the attack. Additionally, it was found that methods based solely on attack time may not be an accurate indication of timbre, since they seem to vary significantly depending on the performer, and performed pitch [10].

3.2 Frequency Domain Approaches

While the amplitude trajectory is clearly important to defining a transient, it does not provide a complete description of the sound. As such, many methods opt to consider the signal in the frequency domain when detecting transients. In particular, a spectral model for transients is useful for applications that aim to label individual audio frames as occurring in a transient or steady state region. These methods are popular in many audio encoding applications, which do not require each note to have only one region of transient. Instead, these methods aim to detect regions of instability in the sound so that the encoding mechanisms can be adjusted accordingly. Since the overtones of a harmonic note tend to roll off relatively quickly at higher frequencies, the energy of these signals will tend to be concentrated in the lower portion of the frequency spectrum. In contrast, transients are often characterized by greater energy in the higher frequencies [1]. As such, regions of rapid change in a sound can be identified by the presence of high frequency energy in spectrum. Detectors that implement this approach often use a small number of bandpass filters and compare the energy in each with the total energy of the sound. If the energy in the highest frequency band is above a certain threshold, then it is classified as a transient the signal is classified as a transient.

Another possible metric that can be used in these applications is a measure of spectral distance between adjacent frames. During the steady state potion of a sound, adjacent frames should share a very similar spectral envelope. In contrast, transient regions are often characterized by rapid changes in the spectrum of the sound. While many methods exist for finding the distance between two spectra, the euclidean distance between the magnitude spectra is commonly used [1].

$$Flux(n) = \sum_{0}^{N-1} |X_k(n)| - |X_k(n-1)|^2 \tag{2}$$

While the metric given here can be used for any type of transient, it is possible to bias this function to onsets by only counting contributions from bins where the energy is increasing [1]. In this case the equation takes the form shown below.

$$Flux(n) = \sum_{0}^{N-1} d_k(n) - |d_k(n)|^2 \tag{3}$$

$$d_k(n) = |X_k(n)| - |X_k(n-1)| \tag{4}$$

Spectral methods of transient detection are often used in audio encoders since they do not make any assumptions about the trajectory of notes or the type of sound being played. When audio is classified on a frame by frame basis in this manner, a far greater number of sounds can be accurately described by the model. In addition to the flexibility of the associated model, these methods remain popular in encoding because they identify transients in regions where the perceptual models used to encode the frequency content of the sound often break down. In a masking based model, signals with large amounts of high frequency would not be well encoded since the masking models account for the higher thresholds of hearing above 10 kHz [2]. Additionally, since they do not compare whole trajectories of the sound, they can be implemented with relatively low latency, which may be desirable.

3.3 Joint Time-Frequency Domain Approaches

As shown above, the description and classification of a transient can be influenced by both the time and frequency domain representations of a sound. As such, there is an increasing trend to use both of these features in segmenting a sound. Almost all of these methods are structured around extending the amplitude trajectory models discussed above to include information from the spectral domain. One of the more basic extensions of the ADSR model is to calculate and compare the energy trajectories separately for different frequency bands. This incorporates the expectation for increased high frequency energy of transients into the calculation, while still giving attention to the full length and time trajectory of the note [15].

While the sub band energy trajectories can be useful in identifying potential transients, it is often desirable to characterize the spectral evolution of a note independently from its amplitude, Although a number of different features can be used to describe these sounds [14], many methods use a combination of the amplitude and spectral centroid [6] [3].

The amplitude trajectory is generally calculated as described above and often smoothed for later analysis. The spectral centroid provides a metric that corresponds to the frequency distribution within a sound and can be calculated using the formula below where $X_k(n)$ corresponds to the magnitude of the spectral bin k at time n. The value of $f(k)$ is generally taken to be the frequency of the kth spectral bin in hertz. [6]

$$CENT = \frac{\sum_k f_k * X_k(n)}{\sum X_k(n)} \tag{5}$$

As can be seen from the equation, the spectral centroid increases when the signal energy is concentrated at higher frequencies, without requiring a threshold to distinguish high and low frequencies like the band pass filters described above. Additionally, the metric is naturally normalized for the loudness of the sound, making it theoretically independent of the associated amplitude trajectory. In order to identify the segmentation points, the first and second derivatives of both the amplitude and centroid are also calculated.

Much like the amplitude trajectory models discussed above, the amplitude centroid trajectory proposed by Hajda divides each note into four distinct segments, which he labels the attack (A), transition (T), sustain (S) and decay(D) [6]. While the sustain and decay portions gener-

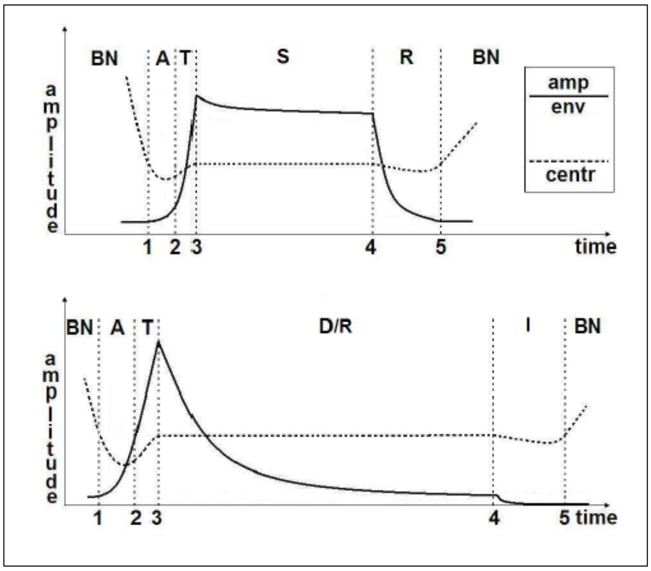

Figure 2. Amplitude and centroid trajectories for sustained and percussive notes as shown in [3]

ally correspond with the analogous portions of the ADSR model, the segmentation of the attack and transition differs substantially. He defines the attack as the portion from the beginning of the note during which the amplitude is increasing and the centroid is decreasing, resulting in a shorter attack length than would occur in an amplitude based model. Despite the high frequency energy that is present in the transient, the centroid can still be expected to be initially decreasing, since the spectral centroid of the background noise will always be one regardless of its loudness [6] The main distinguishing element of Hajda's model is the inclusion of a transition phase between the attack and the sustain, during which both the amplitude and centroid are increasing. [**?**]. The remaining portion of the note is segmented similarly to the ADSR model, however the centroid is assumed to decrease along with the amplitude during the release. In addition to the four main note segments mentioned above, the implementation realized bt Caetano et al also labels segments between notes that it considers background noise (BN) or sections where it thinks the note has been interrupted in some manner (I). Some typical results of this segmentation algorithm are shown in the figure below, and agree reasonably well with where one could expect a transient to be happening.

In addition to incorporating the frequency information of the signal, this method also has the flexibility to deal with percussive as well as sustained tones. Unlike the amplitude trajectory based models which often assume some sustained portion of tone in order to define appropriate thresholds, this method can use the centroid information to easily segment a signal into regions of attack and decay/release without including a sustain.

4. DISCUSSION

Although there are many different methods for detecting and describing transients in audio, they are difficult to effectively compare or rank. While critically important to our perception of sounds and relevant to many applications, a good general definition of an audio transient remains to be found. As a result, the task of transient detection is not driven by the need to agree with a ground truth result as is the case in many other signal detection applications. This means that different applications are able to define different metrics of accuracy depending on their needs, and develop a detection algorithm that is ideally suited to their purpose. Although the various detectors all work well for their given application, the lack of standardization results in a huge number of possible algorithms to choose from. However, some commonalities can be found between the various detection algorithms. In all cases the robustness criteria are very similar, with the goal being to develop a detector that is not influenced by other signal features such as pitch, loudness, performer, tempo, and that works well for a large variety of sounds.

The large number of different transient detectors and segmentation algorithms can also be viewed as an advantage for the researcher in the field. For instance, if the transient is not being explicitly studied (as is the case for many perceptual studies which seek to remove it) a simple amplitude based model may be optimal. While this does not necessarily provide the most accurate segmentation of the note, it can be easily biased to identify a longer transient, ensuring that all the unwanted signal is removed. Also, given the range of low level features that can be used to identify transients, it may be advantageous to choose a detection algorithm that can make use of features that are already being calculated in other parts of the application. For example, in the case of performance analysis where a pitch detector is also being used, much of the spectral and temporal information would already be extracted from the sound, resulting in a minimal extra cost to implement an amplitude centroid type trajectory model. However, some analysis applications that would not otherwise make use of spectral information might be better served by an envelope based detector.

Finally, although this review has focused on segmentation algorithms which assume that a single audio frame should only be categorized with one label, there are cases in which this is not actually the case. For instance, in polyphonic music, a signal could contain the onset of one instrument note as well as the sustained portion of another voice. In this case, it is useful to separate the audio into a sum of transient and steady state signals. In fact, many models choose to further decompose the sound by filtering out both the steady state (harmonic) portion as well as filtered noise component, before identifying the remainder as transient [4]. Under this model, any portion of a sound could contain some of each of these components in varying proportions, with the onsets and offsets of notes being described mostly as a combination of noise and transient. However, this once again introduces a new definition of transient that is connected less with the source sound than with the statistical description of the audio. Much like the earlier definitions used in this article, this definition is also one of exclusion. The transient is something which can not be accurately described in other more well defined ways. Overall, this variation in defining musical transients has led to a variety of highly customized transient detectors which are used in different applications.

5. REFERENCES

[1] J. Bello et al.: "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 5, pp. 1035-1047, 2005.

[2] M. Bosi and R. Goldberg: *Introduction to Digital Audio Coding and Standards*, Kluwer Academic Publishers, Boston, 2003.

[3] M. Caetano, J. Burred, X. Rodet: "Automatic segmentation of the temporal evolution of isolated musical instrument sounds using spectro-temporal cues," *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, Sept 6-10, 2010.

[4] L. Daudet and B. Torresani: "Hybrid representations for audiophonic signal encoding," *Signal Processing*, Vol. 82, No. 11, pp. 1595-1617, 2002.

[5] J. Gordan: 'The perceptual attack time of musical tones," *Journal of the Acoustical Society of America*, Vol. 82, No. 1, pp. 88-105, 1987.

[6] M. Caetano, J. Burred, X. Rodet: "A new model for segmenting the envelope of musical signals: the relative salience of steady state versus attack, revisited," *Proceedings of the Audio Engineering Society Convention 101*, 1996.

[7] H. Jang and J. Park: "Multiresolution sinusoidal model with dynamic segmentation for timescale modification of polyphonic audio signals," *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 2, pp. 254-262, March 2005.

[8] K. Jensen: "Envelope model of isolated musical sounds," *Proceedings of the 2nd COSTG-6 Workshop on Digital Audio Effects (DAFx99)*, NTNU, Trondheim, December 9-11, 1999.

[9] K. Karamitas et al: "A Transient-aware frequency domain audio processor," *Proceedings of the 7th Audio Mostly Conference: A Conference on the Intersection with Sound*, 2012.

[10] D. Luce and M. Clark Jr.: "Durations of attack transients of nonpercussive orchestral instruments," *Journal of the Audio Engineering Society*, Vol. 13, No. 3, pp. 194-199, 1965.

[11] E. Maestre and E. Gomez: "Automatic characterization of dynamics and articulation of monophonic expressive recordings," *Proceedings of the Audio Engineering Society Convention 118*, 2005.

[12] S. Molla and B. Torresani: "Determining local transientness of audio signals," *IEEE Signal Processing Letters*, Vol. 11, No. 7, pp. 625-628, 2004.

[13] G. Peeters: "A Large set of audio features for sound description(similarity and classification) in the CUIDADO project," 2004.

[14] S. Rossignol et al: "Automatic characterization of musical signals: Feature extraction and temporal segmentation," *Journal of New Music Research*, Vol. 28, No. 4, pp. 281-295, 1999.

[15] J. Skowronek and M. McKinney: "Features for audio classification: Percussiveness of sounds," *Intelligent algorithms in ambient and biomedical computing*, Springer Netherlands, pp. 103-118, 2006.

[16] J. Strawn: "Orchestral instruments: Analysis of performed transitions," *Journal of the Audio Engineering Society*, Vol. 34, No. 11, pp. 867-880, 1986.

[17] "Transient, adj. and n.," *OED Online* September 2013. Oxford University Press. 20 November 2013 ¡http://www.oed.com/view/Entry/204789¿.