A high quality sub-band approach to musical transient modification

Markus Zaunschirm
Affiliation1
author1@smcnetwork.org

Joshua Reiss
Affiliation1
author1@smcnetwork.org

Anssi Klapuri
Affiliation3
author3@smcnetwork.org

ABSTRACT

The transient modifier is a type of audio effect which attempts to change the level of the transient parts in a musical signal while leaving the steady state parts unchanged. This paper presents a high performance transient detection and modification algorithm, capable of modifying transients in polyphonic or multi-voiced signals, and capable of modifying both hard (percussive) and soft (non-percussive) transients. The detection and modification are performed in the Fourier domain using a sub band approach. Detection is based on both phase and energy information using an adaptive threshold. Various transient detection functions, based on existing onset detection functions, are compared and real-time implementation issues are discussed. Subjective evaluation in the form of MUSHRA-style (Multi Stimulus test with Hidden reference) listening tests are used to compare the performance of the sub-band approach against other transient modification algorithms.

1. INTRODUCTION

Musical transients are known for holding much of the perceptual information within musical tones. A change in the relation of transient and steady state parts may be used to change the perception of the different notes and instruments. Level changes of transient parts are also used to alter the dynamic range of a music piece. They can inject depth and warmth in a mix and bring out the subtle nuances of the parts. They can also change the perceptual attributes of the mix such as the punchiness or perceived distance of the sources. Unlike other dynamic processors like compression or expansion, which react to the signal level, the transient modifier reacts to the transient content of a signal. The goal of a transient modifier is to modify the identified transient parts while leaving the steady state or non-transient components unchanged, and introducing no or minimal artifacts. The performance of a transient modifier is for the most part determined by the ability to detect transients. Two scenarios are worthy of special attention. The first is when the signal consists of soft transients, generated by non-percussive musical signals. While transient modifiers exist which perform well on a snare track, for instance, it should also be possible to modify the transients of a violin track. The second scenario is based on a polyphonic or multi-voiced signal. A high performance transient modifier should be able to modify all the transients in a signal, even if the signal is comprised of overlapping notes generated by many sources. Finally, the transient modifier should also be easy to use, with minimal manual intervention required, and capable of real-time implementation. These constraints provide the motivation for this work. In Section 2, we provide a definition of the transient in a musical audio signal. In section 2 we summarize and explain different transient detection functions and explore their ability to detect the defined transient parts. The possible frequency-domain modifications are stated in section 4. Based on the results and assumptions of the previous sections we will present an implementation of the entire audio effect in section 5 and discuss its real-time aspects. To evaluate the quality of the suggested implementation we perform several listening tests. The test method and obtained results are summarized in section 6.

2. DEFINITION OF TRANSIENT

In [1] transient parts are loosely defined as short time intervals during which the signal evolves quickly and unpredictably. However, there are many applications related to the detection and modeling of transient phenomena, such as note segmentation for automated music analysis, lossy audio compression, music transcription and onset detection. Each of these may use slightly different definitions. The widely used terms of onsets, attack parts and transient parts are also closely related. Thus it is important to clarify what is meant by a transient in this paper. In general a single note can be segmented in different parts: the attack, decay, sustain and release. According to fig. 1;

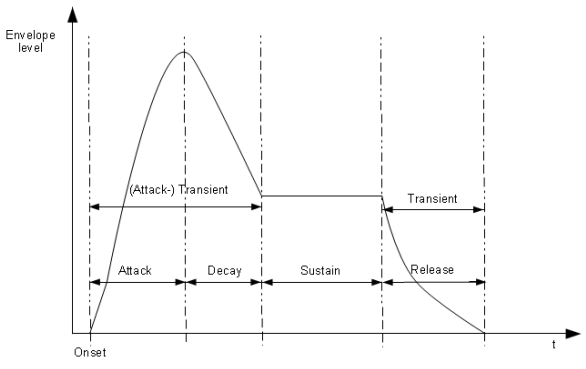

Figure 1. Envelope evolution of a singel note

- the attack is characterized by an increasing amplitude envelope.
- the decay part is the time interval during which the envelope decreases, after reaching the highest amplitude.
- the sustained or stable part exhibits a stable amplitude.
- the release part is the time interval during which the amplitude decreases and the note fades out.
- the onset is a single time instant that should mark the beginning of the temporally extended (attack-) transient, but is at least the earliest time at which the transient can be detected reliably.
- the transients are the time intervals during which the signal characteristics change abruptly and the signal is not stable.

Consequently, a musical signal can be seen as a combination of steady-state and transient parts. This is emphasized by using sinusoidal models [2] for especially lossy audio compression, where most regions of an audio signal are considered as steady-state with respect to the sinusoidal representation and thus can be represented using constant or slowly time-varying parameter trajectories. However, musical signals also contain instances, the transient parts, which can not be described by these steady state conditions. It is stated in [3] that the transient parts are commonly due to abrupt changes in amplitudes, phases, or frequencies, rapid decays in amplitudes and fast transitions in frequencies and amplitudes. Most of these changes occur, when a new note is played and thus are due to energy inputs on the instrument and are thus associated with a note onset. As depicted in fig.1, these (attack-) transients are a combination of the attack and decay part. Fast transitions are considered to be a result of expressive pitch variations such as vibrato passages and often occur in the sustained parts of notes.

2.1 Requirements for transient detection

Since one of the possible applications can be to use transient modification to enhance the impact on the source it is crucial to detect transients associated with the onset (so the attack part). Another possible transient part can be rapid decays in amplitudes, which appear primarily for highly percussive sources such as a snare drum. For these sources a high amplitude or energy increase after the onset is followed by a rapid decay. Using this definition it is reasonable to assume that most of the signal portions of for example a bass drum or a snare are transient parts. To detect these decays negative energy changes also need to be detected. Fast transitions in frequency and amplitudes are mostly related to diverse means of expression such as expressive variations of the pitch or the amplitudes (portato or vibrato passages in music or timbral transitions). These expressive parts should remain unmodified to conserve the artistic content of the modified music signals. This requirement contains a lot of challenging aspects, because in these parts also feature changes in amplitude and phase occur, which can be easily interpreted as transients parts associated with the note onset. Summarized we want to identify the transients associated with the note onset and to certain degree also the decays which can occur after the attack and also after the sustained parts, measure their duration and apply an adequate modification of thess transient portions in terms of softening or hardening the perceived transient.

3. TRANSIENT DETECTION

To define or identify the transient parts of music signals it is necessary to generate detection functions. This can be done by the process of transforming the audio signal into a highly subsampled (or even sample accurate) detection function which manifests the occurrence of transients in the original signal. In general a strong detection function will typically have peaks located at transients, and near zero values elsewhere. The more this is the case, the more robust the detection function can serve as a modification function in terms of modifying just the transient portions and not affecting the steady state portions.

To generate an appropriate transient detection function we analyzed different methods for measuring the transientness of a signal, which are known or used in onset detection, signal coding and speech modeling or speech enhancement techniques. We tried to find the most suitable method for the task in hand in terms of identifying just the previously named transient parts, a low computational effort, a high precision and a suitable signal representation for the following modification without introducing artefacts and keeping the audio quality high.

Since many onset detection approaches use the time varying transientness of a signal to generate an appropriate detection function for the following peak-picking and onset detection it is reasonable to assume that these functions can also be useful for detecting the defined transient parts. Therefore we wanted to take a closer look at the ability of these functions to identify the defined transient parts. We focused on methods, which use a time-frequency representation of the signal and measure the 'transientness' based on changing spectral features. According to [1] these methods can be divided into three main groups:

- Detection based on amplitude or energy information
- Detection based on phase information
- Detection based on combination of phase and energy information.

Further we assumed that a music signal consists of a combination of transient and steady-state parts, and that the transient parts occur as a broadband event and the stable parts as stable pitches with sinusoidal-like components . Therefore we also used a measure of the spectral flatness and the harmonics-to-noise ratio to compute a transient detection function. Other possible approaches are known from signalmodeling. Sinusoidal modeling [2] is

applicable to the analysis, transformation and resynthesis of recorded sound and is also used for a transient-steady state separation. Thereby the audio signal is represented as a sum of sinusoids with slowly varying parameter trajectories. The energy increases of the residual between the original signal and the spectral modeling synthesis (SMS) [4] can be used to detect transient parts and note onsets. Likewise linear prediction is concerned with the estimation of the spectral envelope of signals and the prediction residual can be assumed to be due to transient parts, where the analyzed signal evolves in an unpredictable way.

3.1 Description of the used approaches

In this section we try to give a short description of the used detection functions and to point out their ability to detect transients for the task in hand.

3.1.1 Detection based on Short-Time-Fourier-Transform (STFT)

The STFT of the input signal $x(n)$ is defined as

$$X(n,k) = \sum_{m=\frac{-N}{2}}^{\frac{N}{2}-1} x(nh+m)\omega(m)e^{-\frac{2j\pi mk}{N}} \quad (1)$$

where $\omega(m)$ is an N-point window and and h is the hop size between adjacent windows.

Changes in magnitude: There are several different methods using the energy differences of consecutive STFT frames to measure the transientness of a signal. Due to the assumption that the transients are more noticeable at high frequencies [5] the spectrum can be preferentially weighted towards high frequencies before summing [6]. More general approaches based on magnitude changes of the spectral content formulate the detection function as a distance between consecutive short-time spectra. According to the chosen metric to calculate this distance it is possible to obtain different spectral difference, or spectral flux functions. Commonly the $L1$-norm [6] or the $L2$-norm of the first order difference [7] are used. For generating onset detection functions the differences are rectified in order to count only those frequencies where the energy is increasing. This intended to emphasize onsets rather than offsets. Since transient parts also include decays the transient detection function is computed without rectification. We implemented two transient detection function, which are based on the first order spectral magnitude differences. According to [8] transients can be detected using the transient detection function $T_f(n)$ which is defined as

$$T_{f1}(n) = \sum_{k=0}^{\frac{N}{2}} |(|X(n,k)| - |X(n-1,k)|)|^{\frac{1}{2}} \quad (2)$$

with N as the STFT window length and $X(n,k)$ as the STFT at time instance n for frequency bin k. A refinement takes it cue from psychoacoustics. In [9] it is empirically proven that changes of loudness are judged relative to the overall loudness; this fact is also explained and applied in [10]. A corresponding detection function is:

$$T_{f2}(n) = \frac{\sum_{k=0}^{\frac{N}{2}} |(|X(n,k)| - |X(n-1,k)|)|^{\frac{1}{2}}}{\sum_{k=0}^{\frac{N}{2}} |X(n,k)|} \quad (3)$$

Changes in phase: Approaches which use the phase information of the short-time spectra are generally based on the phase vocoder principle [11]. For the assumption of a steady state or stable sinusoid the phase increment of the $k-th$ frequency bin is defined to be approximately constant:

$$\varphi_k(n) - \varphi_k(n-1) = \varphi_k(n-1) - \varphi_k(n-2) \quad (4)$$

where φ is the unwrapped phase. According to Eq 4 the actual phase deviation between the predicted target and the real phase at time instance n is given by:

$$\Delta\varphi_k(n) = princarg\left[\varphi_k(n) - 2\varphi_k(n-1) + \varphi_k(n-2)\right] \quad (5)$$

where $princarg$ maps the values of the deviation in the range of $[-\pi, \pi]$ and is defined as a modulo operation

$$princarg(\Delta\varphi_k) = mod(\Delta\varphi_k + \pi, -2\pi) + \pi. \quad (6)$$

The defined phase deviation $\Delta\varphi_k(n)$ can now be used to compute transient detection functions. The phase deviation will tend to zero if the phase value is accurately predicted, which implies that the analyzed signal at this time instance consists of stable parts. On the contrary during transient parts $\Delta\varphi_k(n)$ will tend to be large, since the instantaneous frequencies are not well defined for these parts. Basically there are several ways of analyzing the obtained phase deviation of consecutive STFT frames. One approach we used for the generation of a transient detection function is defined as:

$$T_{f3}(n) = \frac{1}{N/2} \sum_{k=0}^{N/2-1} |\Delta\varphi_k(n)| \quad (7)$$

with T_f as the transient detection function at time instance n. Further, in [12] Bello and Sandler presented a method which analyses the instantaneous distribution of the phase deviations across the frequency domain. The kurtosis of the distribution is used to quantify the transientness of the signal (T_{f4}). This approach is justified by the assumption that during steady-state parts the phase deviation for all frequency bins k tends to be zero, thus the distribution is strongly peaked around this value. On the other hand during transient parts the phase deviation for all frequencies which carry transient information tends to be different from zero. Accordingly the distribution during these transient parts is wider and flatter than for stable parts and therefore, the kurtosis of the histogram becomes smaller.

Combination of magnitude and phase information: In [13], both phase and amplitude information work together in the complex domain to quantify the transientness of signal parts. As defined in section 2 for locally steady state regions in the analyzed audio signal the frequency and amplitude should remain constant. Therefore

the magnitude and the phase can be predicted in the complex domain; a target value for each FFT bin can be defined. Each Fourier coefficient is given as a combination of its magnitude and phase $X(n,k) = |X(n,k)| e^{j\varphi_k(n)}$ where $|X(n,k)|$ is the magnitude and φ the phase of the $k-th$ bin at time instance n. The predicted target Fourier coefficient is then defined as $\widehat{X}(n,k) = \widehat{X}(n,k) e^{j\widehat{\varphi}_k(n)}$ where the target magnitude $\widehat{X}(n,k)$ is defined as the magnitude of the previous STFT frame $|X(n-1,k)|$ and the target phase $\widehat{\varphi}_k(n)$ is according to the phase vocoder principle defined by the phase of the previous STFT frames:

$$\widehat{\varphi}_k(n) = princarg\left[2\varphi_k(n-1) - \varphi_k(n-2)\right] \quad (8)$$

with φ_k corresponding to the unwrapped phase of the $k-th$ frequency bin. The transient detection function is then given by the Euclidean distance between the predicted and the actual measured complex Fourier coefficient for each frequency bin k:

$$\begin{aligned} T_{f5_k^2}(n) = &\left[\Re\left(\widehat{X}(n,k)\right) - \Re\left(X(n,k)\right)\right]^2 \\ &+ \left[\Im\left(\widehat{X}(n,k)\right) - \Im\left(X(n,k)\right)\right]^2 \end{aligned} \quad (9)$$

A frame-by-frame detection function is defined by:

$$T_{f5}(n) = \sum_{k=1}^{N} \sqrt{T_{fk}^2(n)}. \quad (10)$$

Spectral flatness measure: Generally the spectral flatness measure (SFM) is used to determine the amount of randomness that is present in a signal [14]. Therefore it describes the characteristics of a spectrum; to be precise, whether a signal has a noisy or 'sinusoidal' spectrum. So it is also called 'tonality coefficient' and is used to quantify how much tone-like a sound is, as opposed to being noise like. The received values of a SFM are bounded between $0 \leq SFM \leq 1$. Thereby a low spectral flatness indicates a structured or non- random process with the spectral power concentrated in a very small region of the spectrum. In contrast a high spectral flatness corresponds to a random signal; accordingly the spectrum has a similar amount of power in all spectral bands. So according to the definitions of transients (random, broadband event) we used the spectral flatness measure as a transient metric. Which is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum

$$T_{f6}(n) = \frac{\sqrt[N]{\prod_{k=1}^{N} P(n,k)}}{\frac{1}{N}\sum_{k=1}^{N} P(n,k)}. \quad (11)$$

3.1.2 Detection based on harmonics to noise ratio

The Harmonics to Noise Ratio (HNR) is an acoustic parameter for the objective description of voice quality. Therefore the harmonic structure of an audio signal is examined and then used to determine the amount of (non-harmonic) noise present in addition to the harmonic parts [15].For the estimation of the HNR of signals, cepstral analysis is used. The cepstrum is produced for a windowed segment of the audio signal. Cepstral peaks at the fundamental period and multiples thereof are considered to be due to the harmonic signal. These detected rahmonics are zeroed and the resulting liftered cepstrum is inverse Fourier transformed to provide a rough estimate of the additive noise. To maintain the adequate noise spectrum a baseline correction procedure has to be applied to the first estimate (because the first estimate is too high). After the correction the logarithm of the summed noise energy is subtracted from the logarithm of the summed original audio signal energy in order to provide the HNR estimate

$$T_{f7}(n) = 10log\frac{\sum_{k=1}^{\frac{N+2}{2}} |X(n,k)|^2}{\sum_{k=1}^{\frac{N+2}{2}} |N(n,k)|^2} \quad (12)$$

where $X(n,k)$ denotes the kth frequency bin of original signal and $N(n,k)$ the corresponding estimated noise spectrum bin. In decibels equation 12, yields to a simple subtraction and the Tf is defined as $T_{f7}(n) = X(n)_{[dB]} - N(n)_{[dB]}$.

According to this definition the transient detection function should yield high values for stable parts, where the signal is assumed to consist of fundamental frequencies and their harmonic components. In order to compare the results to the previous approaches the results were reflected to have a maximum for transient parts.

3.1.3 Based on signal modeling

Linear Prediction Generally, linear prediction is concerned with the estimation of the spectral envelope of speech signals and is based on autoregressive (AR) modeling. An overview is given in [16]. Since AR modeling is a process which is often used to model and predict various types of natural phenomena we used it for the modeling of musical signals. A description of the input signal $x(n)$ using an AR process can be written as:

$$x(n) = \sum_{m=1}^{p} c_m x(n-m) + u(n) \quad (13)$$

where p is the model order, c_m are the prediction parameters of the model and u is a noise-like statistically independent signal. Accordingly, the modeled signal can be defined by the excitation $u(n)$ and a sum of the previous samples $x(n-m)$. Further the predicted signal can then be defined as

$$\widehat{x}(n) = \sum_{m=1}^{p} c_m x(n-m). \quad (14)$$

When the signal can be modeled by linear prediction, it is assumed to be stationary or quasi-stationary. However, at the transient parts this stationarity assumption fails to hold, which leads to a significant increase of the prediction error. In order to compute a transient detection function the residual signal $x(n) - \widehat{x}(n)$ (T_{f8}) was assumed to be a sequence of peaks at the locations of transients; the parts which can not be predicted well.

Sinusoidal Modeling As stated before Sinusoidal modeling [2] is applicable to the analysis, transformation and

resynthesis of recorded sound. Thereby the audio signal is represented as a sum of sinusoids with slowly varying parameter trajectories:

$$\widehat{x}(n) = \sum_{k=1}^{p} A_k(n) cos\left(\varphi_k + \sum_{s=0}^{n-1} \omega_k(s)\right) \qquad (15)$$

where $A_k(n)$ is the amplitude of the $k-th$ sinusoid, φ_k is the phase and $\omega_k(n)$ is the frequency at time n, with $n = 1....N$ and N refers to the total number of analyzed blocks. As for the linear prediction, the signal parts which can be described meaningfuly by a sinusoidal model are assumed to be stationary. Transient parts violate these steady state condition and therefore, the energy increases of the residual between the original signal and the spectral modeling synthesis (SMS) can be used to detect transient parts $T_{f9}(n) = (x(n) - \widehat{x}\,(n))^2$ [4].

3.2 Results

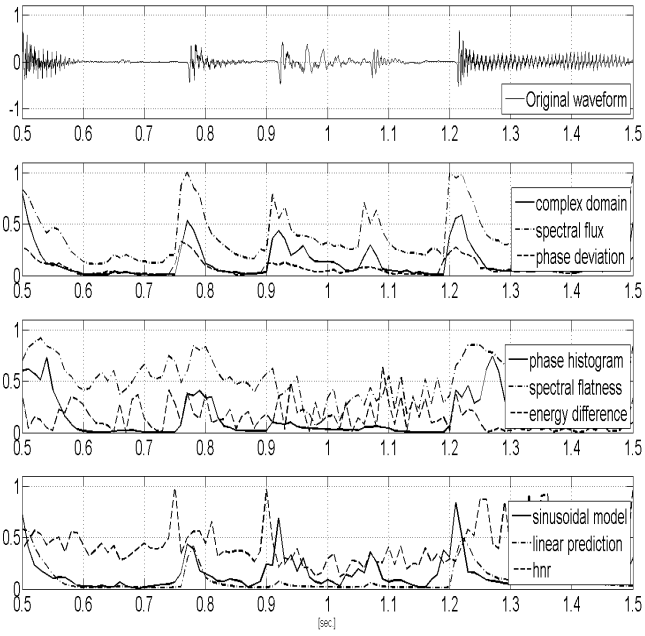

Figure 2. Transient detection functions of a slapped bass sample

In this section we will summarize the shortcomings of the different evaluated approaches to detect transient behavior in musical signals and show figures to illustrate the behavior of the detection functions for two music samples (they are not intended to serve as an evaluation). As can be seen in fig.2 and 3 the results for the spectral flatness and harmonics-to-noise measures do not come up to the requirements of an adequate detection function. Although, the functions exhibit peaks at the location of the assumed transient regions they also show lots of spurious peaks located elsewhere and accordingly, these approaches are excluded from the pool of the possibly robust detection functions. The detection functions which make use of the residual signal between the modeled and the original signal do

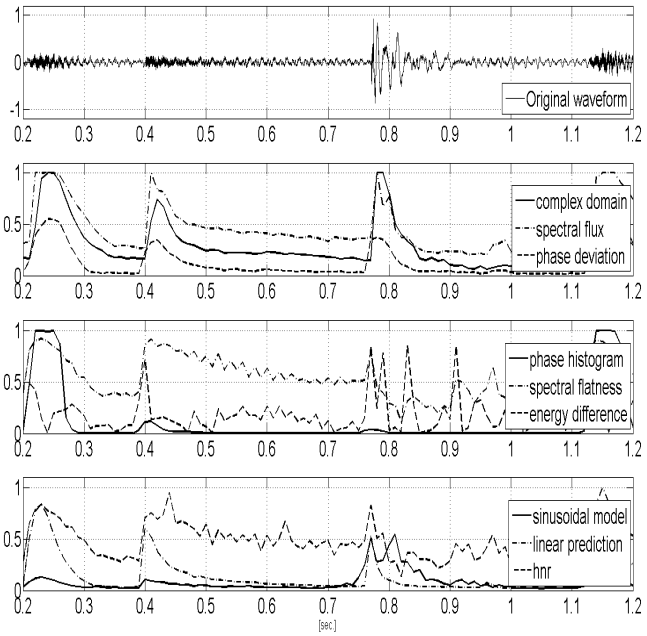

Figure 3. Transient detection functions of a complex mix (pop)

seem promising. The differences between assumed transient and stable parts are more distinctive when compared to the SFM and the HNR, but do not show a clear improvement when compared to the detection functions that are based on the differences between STFT frames.

Since it is possible to modify the signal in the frequency domain with perfect reconstruction, we decided to use one of the more straightforward approaches that work in the STFT domain. In sec. 2 we stated that transient regions are generally due to abrupt changes of amplitude, phase and frequencies. The approaches using the magnitude information are able to detect the rapid amplitude changes, but the approaches using the phase information possibly just detect abrupt changes of phases and frequencies. Since the complex domain approach defined by Duxbury [13] offers a combination of both the phase and magnitude information we focused on this method for the further computations. Further confirmation of this approach are given in [1] and [12]. It is stated that the effectiveness of energy based algorithms decreases for nonpronounced transients in a signal. Further algorithms based on the phase deviation perform well when detecting the transients of pitched sounds, but poor for complex mixes and purely percussive sounds, since these approaches are susceptible to phase distortion. Accordingly, the combination of the phase and magnitude information should enable an equal detection of pronounced (percussive) and non-pronounced (pitched) transients in complex mixes and also for single instruments.

As depicted in fig.2 and 3 (upper plot) the combination of phase and magnitude information yields a smoother, less noisy detection function; *compared to spectral flux and*

phase deviation.

3.3 Transient duration

In general we assume that the computed detection function exhibits higher values for transient parts than for stable parts. Since the detection function is bounded between $[0, 1]$ it would be possible to set a fixed threshold at 0.5, where values above the threshold are defined as transient parts and successive values above the threshold define the time duration of a transient region. Since the absolute values of the detection function are highly dependent on the temporary energy and frequency density a fixed threshold would not give satisfying results. Accordingly the most relevant information to estimate the transient durations is embedded in the relative difference between the values for assumed stable and transient parts. Therefore, a transient duration is modeled by the temporally extended peaks in the detection function. It is stated in [17] that the median filtered detection function can be used to obtain an adaptive threshold for the detection of impulsive noise in music signals, which is a comparable task. The adaptive threshold is defined as

$$\vartheta(n) = a \cdot median(w(n)) \tag{16}$$

with a as a scaling factor and $w(n)$ as a sequence of the detection function given by

$$w(n) = \{Tf(n-i), ...Tf(n)...., Tf(n+i)\} \tag{17}$$

The corresponding length of the median filter is $2i + 1$. In order to prevent the threshold from rising at the position of a peak the length of the median filter has to be set longer than the assumed duration of the peak in the transient detection function. So the length of the median filter is set to a value, which is dependent on the assumed maximum transient duration and the time resolution of the detection function. Summarized the detection of transient parts is

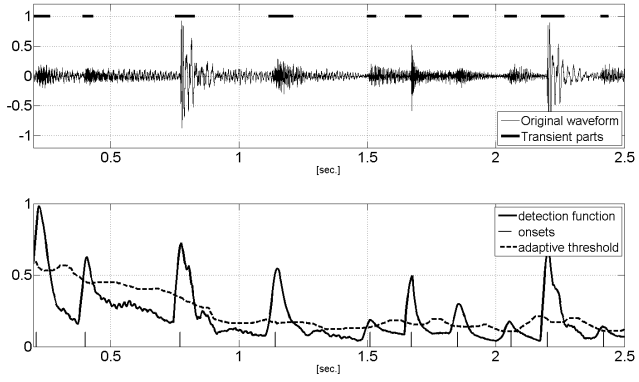

Figure 4. Estimation of transient parts using the complex domain detection function and an adaptive threshold.

carried out by forming a detection function with peaks indicating the transientness at every time instance. These peaks are judged to represent transient regions by an adaptive threshold curve. The estimated transient regions for a Pop sample are shown in fig. 3

4. TRANSIENT MODIFICATION

Let us consider a complex Fourier coefficient for the $k-th$ frequency bin at time instance n as a combination of its magnitude and phase:

$$X_k(n) = |X_k(n)|\, e^{j\varphi_k(n)}. \tag{18}$$

The aim is to implement a dynamic processor that reacts related to the transientness of the audio input and not to its level. If the audio signal at time instance n is defined as transient the magnitude of all bins at this corresponding time instance should be modified according to the detected relative transientness. Let us define this relative transientness as the difference between the transient value $T_f(n)$ and the threshold value $\vartheta(n)$:

$$\Delta\tau(n) = T_f(n) - \vartheta(n). \tag{19}$$

According to [8] the amount of modification should be variable and should be a function of user inputs. Therefore let us define the actual modification value $G(n)$ by the user input g and the detected relative transientness $\Delta\tau(n)$:

$$G(n) = F(\Delta\tau(n), g) \tag{20}$$

High $\Delta\tau(n)$ indicates transient regions more reliably than a small $\Delta\tau(n)$. In order to minimize the effect of false positives, the modification or the amount of modification should not be carried out equally for every detected transient. The modification should be dependent on the difference between the transient value and the threshold value and a smaller difference also result in less modification. The used tuning functions for an amplification scenario and $g = 1.2$ are shown in fig.5. The linear function maps the values of the relative transientness between $[0, 0.5]$ linearly onto modification values between $[1, g]$, whereas the soft mapping functions maps small transientness values to small modification values.

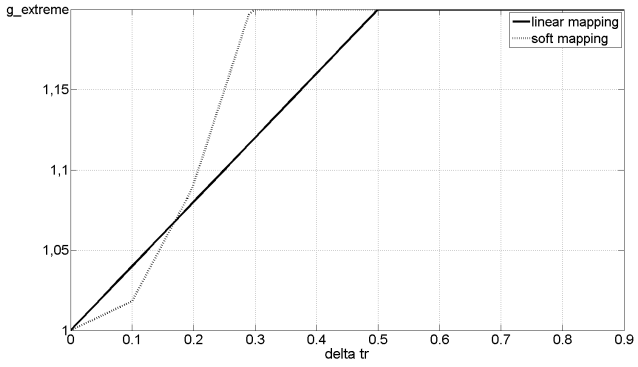

Figure 5. Modification values as a function of the relative transientness and gain parameter g for a transient amplification scenario

The frequency-domain transient-based modification was introduced by Goodwin in [8]. Thereby the modified complex Fourier coefficient $\widetilde{X}_k(n)$ is obtained by a function of the actual modification value $G(n)$ and the original Fourier

coefficient $X_k(n)$. Goodwin distinguished between two modification schemes. The linear modification, where the magnitude of the original coefficient is multiplied by the corresponding modification gain value,

$$\left|\widetilde{X}_k(n)\right| = |X_k(n)|\, G(n), \tag{21}$$

and the nonlinear modification, where we use a slightly different computation:

$$\left|\widetilde{X}_k(n)\right| = (|X_k(n)| + 1)^{G(n)^2} - 1. \tag{22}$$

[8] stated that the nonlinear modification yields more natural sounding modifications, and that dominant spectral components are more affected by the nonlinearity, since in a complex mix transient parts are assumed to be dominant over stable parts. Since according to [5] transients are more noticeable at high frequencies and these frequencies usually carry less energy than low frequencies, we also defined a nonlinear modification scheme that affects spectral components with less energy more.

$$\left|\widetilde{X}_k(n)\right| = |X_k(n)|^{\frac{1}{G(n)^{\frac{1}{3}}}}. \tag{23}$$

All of these modification processes preserve the phase of the original audio signal. How these different modification schemes affect the magnitudes of the modified signal can be seen in fig.6. As stated, all previous modification

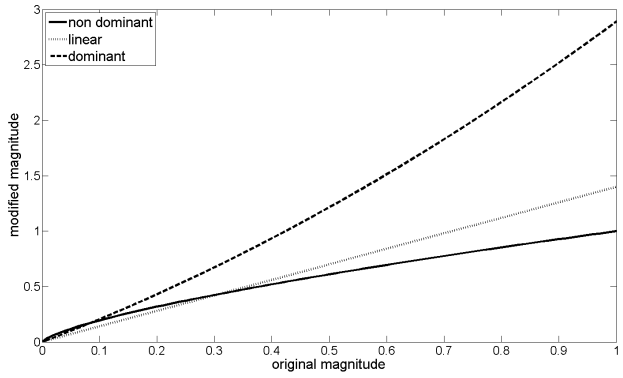

Figure 6. Modifications of output magnitude for a fixed modification value $G = 1.4$: linear(eq. 21), dominant (eq. 22) and nondominant(eq. 23)

schemes change the magnitude of a complex Fourier coefficient and not its phase. However, according to [18] perception of transient parts is mainly dependent on phase relationships during these parts. Therfore, we also tested a modification scheme, where just the phases of each bin are affected by the modification:

$$\widetilde{\varphi}_k(n) = \varphi_k(n)^{G(n)^2}. \tag{24}$$

This kind of modification is just applicable for transient suppression. Especially the attack-transients are perceived softer and sound blurred.

5. SUB-BAND PROCESSING

If we consider a complex mix, where different instruments play different notes at different time instances, we can assume that the resulting spectrum is a mixture of transient and nontransient events. Therefore, it is also reasonable to assume that at a specific time instance the note of one instrument is in a transient part whereas the note of another instrument is in its stable part. Accordingly, if the transient part at this specific time instance is detected, the modification of the whole spectrum would also lead to a modification of the overlaid stable part. Since the set goal is to modify just the transient parts a modification of the whole signal is not suitable. Further, when considered a single instrument, the attack times for different frequencies (harmonics) are assumed to have different durations. It is also stated in [10] that especially low parts of a note may take some time to come to the point where their amplitude is maximally rising, which leads to an incorrect cross-band association with the higher frequencies. To overcome these problems we decided to implement the actual transient-based frequency domain modification as a subband approach. If for example a stable note with a fundamental frequency of 400 Hz is played and another event with a fundamental frequency of 2000 Hz starts at the same time, the transient modification of the 2000 Hz event should not lead to a modification of the stable note at 400 Hz. For subband transient detection and modification, each subband in each window is characterized as being steady state or transient, as opposed to a single, transient function implementation. Since some onset detection studies have found it useful to independently analyze information across different frequency bands, it is also reasonable to measure the transient behavior for different subbands. This approach is also applicable in order to adjust the detection to the human auditory system by treating frequency bands seperately. Another advantage of a subband approach is that it is possible to detect amplitude and phase changes in different bands. For example in polyphonic music it could occur that smaller amplitude changes are not noticeable because of stable parts with higher energy masking the transient. By using different subbands this effect can be reduced. Another advantage is that the modification can be done flexibly, since every subband has its own transient detection and weighting function. The general scheme of the system is shown in fig. 7. The different stages of the detection and

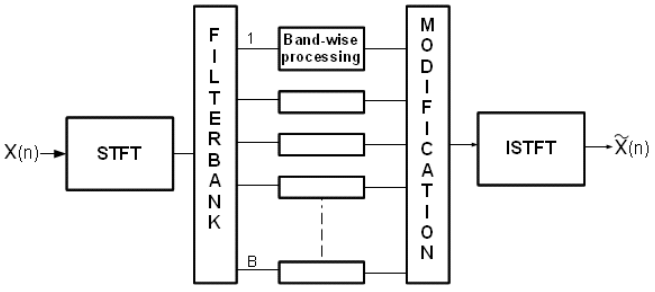

Figure 7. System overview

modification process are discussed next.

Audio input: The input signal can be a monoaural or stereo signal at any sampling rate. In order to compare the results with the original file, the maximum level of the input signal is reduced to leave headroom for the modification.

STFT: The time resolution has crucial influence on the detection and modification. It is known that transient portions have generally a short duration, which is assumed to be between 30 and 130ms. Therefore, a good time resolution is needed. On the other hand the frequency resolution should also be high to get better results. The used STFT has a framesize of $N = 2048$ and a hopsize of $L = 512$ samples (this equals a time resolution of 11.6ms at fs=44100 Hz).

Filterbank: The resulting spectrogram can be split into several non-overlapping sub-bands according to the logarithmic scale between the frequency range of 20Hz to 19845Hz. We used $B = 6$ subbands which leads to roughly a bandwidth of 1.5 octaves per band. The subband processing is done as stated in the previous sections for every subband and is depicted in fig.8.

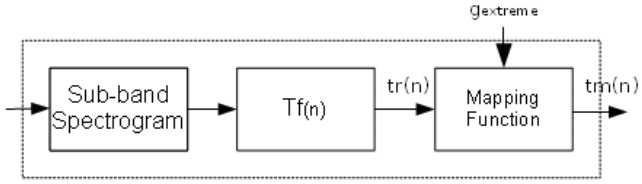

Figure 8. Processing stage

Summarized, the transient detection functions $T_f(n)$ are generated according to equations 9 and 10 and are adapted for the human auditory system by weighting it with the predominant energy at each time instance

$$T_f(n) = \frac{T_f(n)}{\sum_{k=0}^{\frac{N}{2}-1} |X(n-4,k)|} \tag{25}$$

. This is based on the fact that intensity differences are perceived related to the overall intensity. Further the absolute transient values are bounded between $[0, 1]$ using a simple peak follower

$$T_f(n) = \frac{T_f(n)}{\gamma(n)} \tag{26}$$

with $\gamma(n) = max(T_f(0).....T_f(n))$. The adaptive thresholds for each sub band are obtained according to equations 16 and 17, with a scaling factor $a = 1.2$ and a filter length $i = 14$. Next the computed relative transientness values $\Delta\tau(n)$ are mapped according to the soft tuning characteristics to obtain the transient modification values $G(n)$. Figure 9 shows the transient modification values for a suppression scenario and a $g = 0.6$. The actual modification can be performed as stated in equations 21, 22 and 23.

User settings: In order to allow the user to adjust the behaviour of the detection and modification we implemented the global parameters as given in table 1.

The most important setting is the amount of modification g which can be set independently for each sub band.

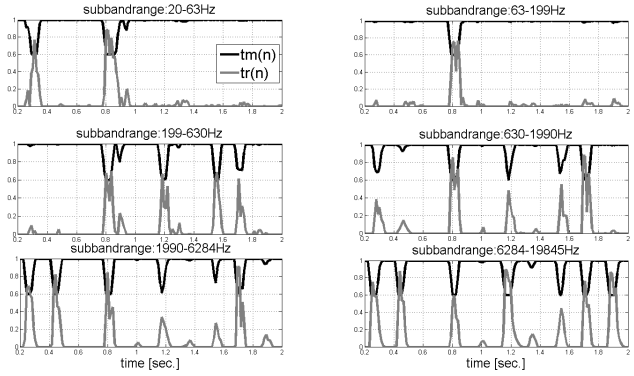

Figure 9. Relative transientness and resulting modification values for a segment of a pop sample

Parameter	Suppression	Amplification
g	0.4-0.99	1.01-1.8
$mod_{min}\,[dB]$	0-4	0-6
a	1-1.5	1-1.5
filter length i	10-40	10-40

Table 1. Global parameters used for transient modification

The factor $mod_{min}\,[dB]$ defines the lowest amount of modification that is realized. For example if the parameter $mod_{min} = 4\,[dB]$, no modification that results in level increase or decrease of less than $4\,[dB]$ will be applied. The parameters a and i can be used to change the behavior of the median filter. a changes the overall level of the threshold and the parameter i the filter length; for higher a, just stronger transients will be detected. Figure 10 shows the original, modified and residual signal for a suppression scenario. Demonstrations of the modified samples using

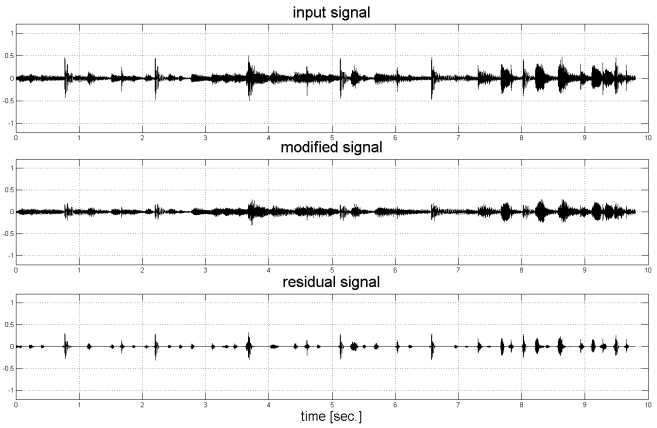

Figure 10. Audio input, resulting audio output and residual signal for transient suppression of a complex mix; $g = 0.6$

all modification schemes (section 4) are available at (link to c4dm page).

Real-time aspects In general it is possible to implement the detection and modification in real-time. The in-

troduced latency is mainly determined by the STFT, the needed overlap and add for the corresponding IFFT (Inverse-Fourier-Transform) and the design of the median filter for the threshold generation. According to equations 16 and 17 the median filtering requires a lookahead time of i frames. To reduce the overall latency a median filtering approach without lookahead can be used. This scheme can reduce the latency significant, but also affects the detection and modification behavior of the entire audio effect. The possible impact on the detected transients and their durations are depicted in fig. 11. It can be seen that the two threshold schemes are comparable and also detect the same transient parts but with slightly different durations.

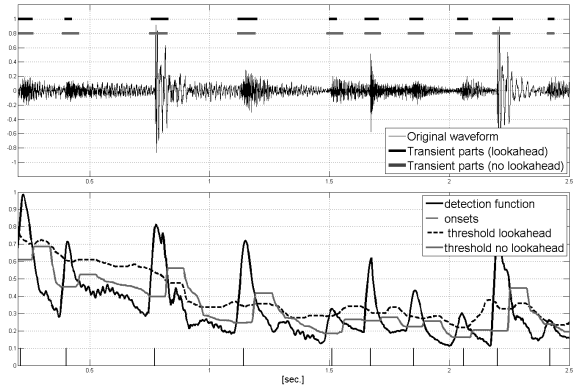

Figure 11. Comparison of detected transient durations for different threshold functions.

6. LISTENING TEST AND RESULTS

To investigate the performance of the implemented transient detection and modification approache we set up a listening test. The general goal of the listening test was to compare the implemented sub-band approach with adaptive threshold (explained in 5), against a single band approach with adaptive threshold (based on the same scheme as the subband approach, but applied to the whole frequency rage) , an all-band approach (every frequency bin is treated as different subband) , a single band approach with fixed threshold (basically our implementation of the approach presented in [8]) and a consumer VST-plugin. Primarily, we wanted to find out if the sub band approach is an improvement to the fixed threshold approach and how it compares to an existing tool. Further, we wanted to find out if there is a difference in the results for different subbands and so we also tested the extreme settings (just one band, and $N/2$ bands, with N referring to number of frequency bins in the spectral domain).

We first generated a set of audio stimuli for use in the listening test. The reference signals were chosen from different genres. In order to get a diversified sample pool we used percussive, pitched-percussive and non-percussive sounds, as well as monophonic and polyphonic samples. All samples were modified using the defined different approaches. The parameters were set as constant as possible to ensure a fair non-discriminatory comparison.

In general the listening test should verify the perceptibility of the implemented effect and evaluate the quality of the resulting audio output for the different approaches. Further we tested the change of the perceptual attribute punchiness-forcefulness and the perception of distance for different amounts of modification. Our research question here is the extent to which the modification is audible and to what extent it impedes the perceived quality of the produced audio output. The titles or questions asked in the different experiments are listed below:

- Rate the following samples according to perceived transient suppression.
- Rate the following samples according to the ability to modify transients from all sources.
- Rate the samples according to the punchiness.
- Rate the samples according to the perceived distance.
- Rate the samples according to the ability to amplify the transients while not affecting the steady state portions.
- Rate the samples according to their modification quality.

The full instructions are presented at (link to the c4dm page).

6.1 Method

Since transient modification affects mainly the level of transients with respect to the steady state portions (changes also the dynamic range of the input audio), we did not normalize the resulting output to equal loudness or a maximum peak-level. This ensured the best possible comparison in terms of rating the change of the overall-signal level, the amount of amplification and suppression during transient parts and the general change compared to the reference. In order to avoid clipping the reference signals were normalized to a maximum peak value of 0.5. For the generation of the modified samples we set the modification parameters in order to achieve a similar maximum modification for all approaches under test. Since the consumer VST plugin offers a variety of different parameter settings we set the input gain to zero, the overshoot value near the middle of the scale (about 8 ms) and the amount of modification (ratio9 was set low to avoid clipping for the whole range of output signals. The generated audio excerpts are published online. Tests were performed in a framework related to the 'Multi Stimulus test with Hidden reference and anchor' (MUSHRA) standard [19]. Generally, participants in MUSHRA tests are presented with sets of processed audio excerpts and asked to rate their basic audio

quality compared to an unprocessed reference audio excerpt. Usually, each excerpt set includes excerpts produced by the investigated applications, the unprocessed audio as hidden reference and a 3.5kHz low-pass filtered version of the excerpt as a low-quality anchor. Since this method is designed for the subjective assessment of intermediate audio quality and not for the evaluation of an audio effect, the 3.5kHz low-pass filtered original is not used as anchor. For experiments set up for the evaluation of the ability to amplify or suppress the transients parts, the hidden reference is used as anchor. This is reasonable due to the fact that the subjects have to rate the perceptual change in relation to the original. But even for the evaluation of the modification quality the suggested anchor is not appropriate. We used anchors showing similar types of impairments as the output of the single-band and fixed threshold approach. So in general the test may be considered as a mixture of a MUSHRA-test and a semantic differential or rating. The test included 8 experiments, and the order of the experiments and of the excerpts within each trial was randomized. Each experiment contained 6 signals to be graded, between 3 to 15 s long. The subjects could listen to the signals in any order, any number of times. The grading scale was continuous from excellent to bad where a grade of 0 corresponds to the bottom of the bad category, while a grade of 100 corresponds to the top of the excellent category (the interface is shown in fig. 12. The instructions for each experiment were given on an additional sheet.

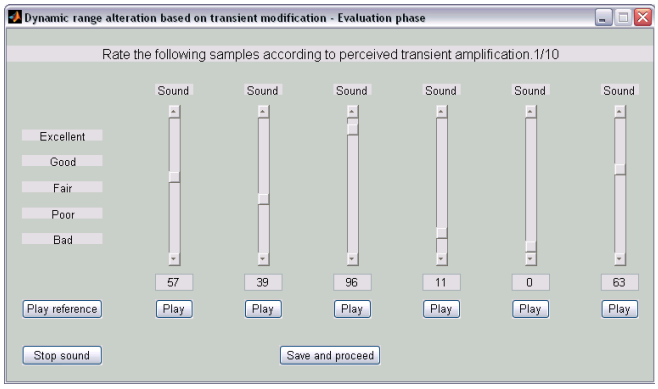

Figure 12. Interface of the listening test.

Participants: We recruited 13 experienced music listeners (9 men and 4 women aged between 24 and 34). Tests were performed using headphones and took around 30-40 minutes in total to complete, including the initial training phase and the actual evaluation phase. For the prescreening we made sure that all tested subjects had normal hearing. For post-screening we performed a combination of numerical test and manual inspection. We calculated the Pearson's r and Spearsman's φ for each participant of their gradings with the median of the gradings provided by all participants. Sets of gradings with a low correlation were considered to be possible outliers and accordingly, inspected manually. Further sets of gradings in which the participants did not rate the hidden reference consistently were also considered to be outliers and inspected manually.(For rating the quality the hidden reference should be rated 100; for rating the perceived suppression/amplification and changed punch the hidden reference should be rated 0). However, since it is stated in [20] that participants tend to treat MUSHRA-related taks to some extent as a ranking task and therefore penalize the hidden reference slightly if they misidentify other signals as the signal with the highest quality (or in our test misidentify other signals as the signal with no transient modification) we did not reject them automatically. Dependent on the correlation values and the rating of the reference we identified outliers for each experiment as given in table 2:

experiment	outliers
modification of all sources	1
transient suppression	1
increased punch	1
not affecting steady state	0
perceived quality	3

Table 2. Identified outliers for each experiment

Analyzing listening test data: In the MUSHRAM standard [19] it is recommended to analyze the results using a mean score and the associated confidence interval according to a t-distribution for each experiment. In order to determine whether the differences between the results of different conditions were significant we performed paired sample t-tests with a significance level of 0.01 and 0.05. We applied the t-test in the original domain and also for the logistic transformed data, since results are assumed to be more meaningful in this representation [21]. The results for the t-test are included in the interpretation of the results.

6.2 Results

The results for all experiments are presented in this section. Each plot shows from left to right the ratings for the hidden reference (*ref*), the sub band approach using 6 subbands (*bands*), the all bins approach were each bin is treated as different sub band (*all bins*), the one band approach with adaptive threshold (*one band*), the one band approach with fixed threshold (*fixed th.*) and the samples generated using a consumer vst-plugin. **Ability to modify transient parts from all sources:** The aim of this experiment was to find out which approach works best on the modification of a complex mix, in terms of changing the transient level of all sources, not just transient parts of highly percussive or louder sources. The results are shown in 13.

According to the mean value of the ratings, the subband approach performs best for the modification of transients from all sources. The results of the t-test for this experiment implies that there is no significant difference between the results, except between the results of the subband and consumer ratings. **Perceived transient suppression:** For this experiment we used a slap bass sample as reference signal. The participants were asked to rate ac-

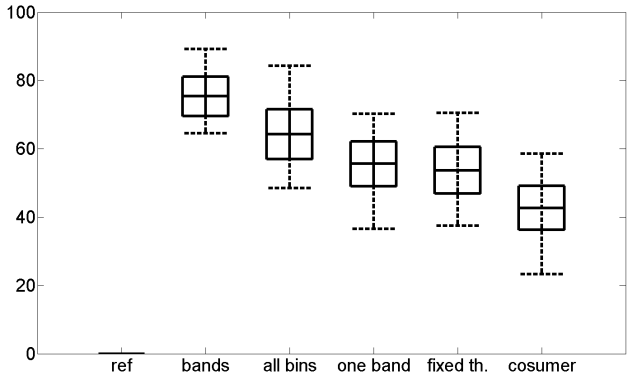

Figure 13. Results from the listening test (*Ability to modify transients from all sources*), showing mean and 95% confidence interval according to a t distribution and whiskers extending to the 25 and 75 percentiles.

cording to the perceived transient suppression; e.g. if the bass seems to be played softer. It can be seen in fig.14

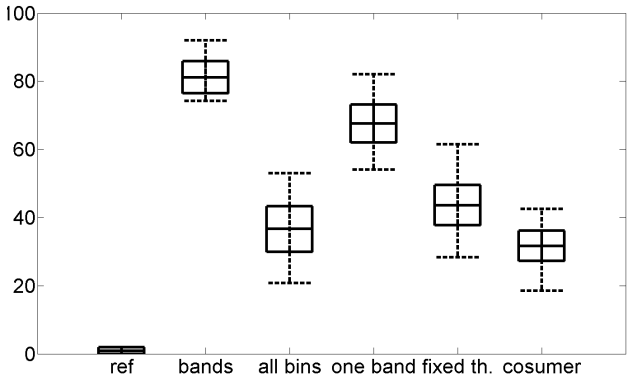

Figure 14. Results from the listening test (*Perceived transient suppression*), showing mean and 95% confidence interval according to a t distribution and whiskers extending to the 25 and 75 percentiles.

that the bands and one band approach perform significantly better than the other approaches. **Increased punch:** The reference signal for this trial was a drum sample. All modified signals were rated higher than the hidden reference but did not exhibit significant differences to each other; therefore results are not depicted. Accordingly a change of the perceptual attribute punch was audible for all approaches. **Perceived distance:** The participants were asked to rate samples according to the perceived distance; 100 very near and 0 far away. As reference we used a conga sample. For the generation of the samples we used the sub band approach and changed the amount of amplification and suppression. According, to the results presented in fig. 15 it is possible to change the perceived microphone-source distance by changing the relation between transient and steady state parts. **Modification of transients while not affecting steady state portions:** As reference signals we used a drum sample with added stable sinusoids and a bowed

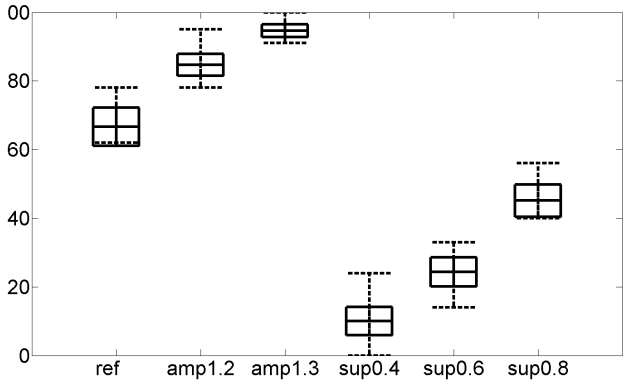

Figure 15. Results from the listening test (*Perceived distance*), showing mean and 95% confidence interval according to a t distribution and whiskers extending to the 25 and 75 percentiles; amplification(amp) or suppression(sup) with the following values indicate the amount of modification.

string. The participants were asked to rate to what extent steady state portions of the signal were modified. This experiment was intended to spot which approach is able to detect transient parts and their duration most meaningful. Further it should also verify the ability of the bands approach to not affect steady state portions in the presence of a transient part;if they are located in different sub bands. The sub band approach and the consumer tool equally out-

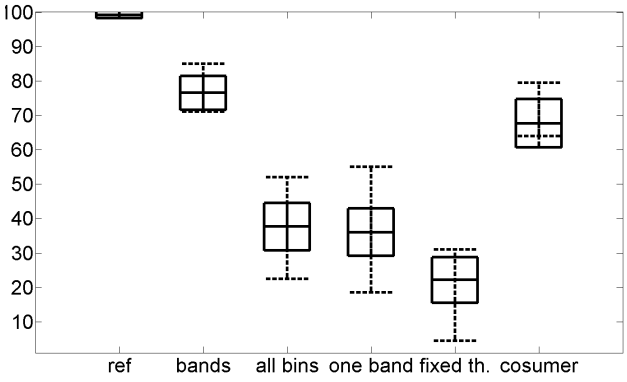

Figure 16. Results from the listening test (*Modification of transients while not affecting steady state portions*), showing mean and 95% confidence interval according to a t distribution and whiskers extending to the 25 and 75 percentiles.

perform the other three approaches significant. However the difference between the two best performing approaches is not significant enough to spot a difference; see fig. 16.

Perceived quality: As reference we used a complex mix (drums,piano,bass) and a polyphonic sample (guitar,drums). The aim was to find out which approaches impair the perceived quality and how the quality is rated compared to the hidden reference. The samples generated using the sub-

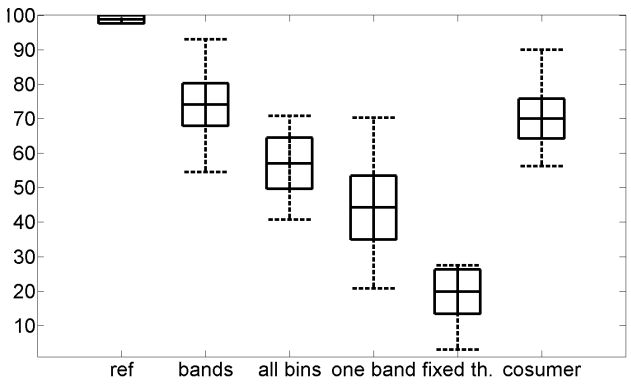

Figure 17. Results from the listening test (*Perceived quality*), showing mean and 95% confidence interval according to a t distribution and whiskers extending to the 25 and 75 percentiles.

band approach and consumer tool are significantly rated as having the least quality impairment. It also can be seen in fig. 17 that the fixed threshold approach introduces the most unwanted effects.

7. CONCLUSION

In this paper, a new, high performance transient modifier was developed and evaluated. First, different onset detection methods were compared in order to determine a suitable transient detection function. Our approach used a complex domain onset detection function with sub bands and a short time Fourier transform in order to modify only those bands and time intervals with significant transient behavior. An adaptive threshold was used to adapt to changing dynamics and signal levels, and transient modification was based both on the frequency range of the sub band and the relative level of the transient detection function for that sub band. Real-time versions implemented with either a look ahead (increased latency) or median filtering (inaccuracies in transient duration measurement) were discussed. MUSHRA-style listening tests were performed in order to compare this approach against other approaches for several performance measures. It was shown that the sub-band, adaptive threshold approach outperformed adaptive and fixed threshold approaches. Contrary to our expectations, the sub band approach also generally outperformed an approach where all frequency bins may be modified independently. A possible explanation for this could be that constant time resolution as a function of frequency led to poor results for high frequencies. It may be possible to improve this performance using a multi-resolution STFT or a constant-Q transform. Although real-time implementations were discussed and analysed, they were not evaluated, and subjective evaluation would be necessary to determine whether the real-time methods would lead to perceptually worse performance. Finally, the modification was restricted to amplification or suppression of the transients. Our approach should allow for more creative forms of transient modification.

A. REFERENCES

[1] J. Bello, L. Daudet, S. Abdullah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, September 2005.

[2] X. Serra and J. Smith, "Spectral Modeling Synthesis: a Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[3] H. Thornburg, *Detection and Modeling of transient audio signals with prior information.* PhD thesis, Stanford university, 2005.

[4] S. N. Levine, T. S. Verma, and J. O. S. Iii, "Multiresolution sinusoidal modeling for wideband audio with modifications," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, (Seatlle), 1998.

[5] X. Rodet and F.Jaillet, "Detection and modeling of fast attack transients," *Int. Computer Music Conf.*, pp. 30–33, 2001.

[6] P. Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals*. PhD thesis, University of Bristol, UK, 1996.

[7] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Digital Audio Effects Conf. (DAFX,'02)*.

[8] M. M.Goodwin and C. Avendano, "Frequency-domain algorithms for audio signal enhancement based on transient modification," *J. Audio Eng. Soc.*, vol. 54, September 2006.

[9] B. Moore, B. Glasberg, and T. Bear, "A model for the prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–239, April 1997.

[10] A.Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-99)*, (Phoenix), pp. 115–118, 1999.

[11] R. Fischman, "The phase vocoder: theory and practice," *Org. Sound*, vol. 2, no. 2, pp. 127–145, 1997.

[12] P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proceedings of IEEE ICASSP*, 2003.

[13] C. Duxbury, J. P. Bello, M. Davies, M. Sandler, and M. S, "Complex domain onset detection for musical signals," in *In Proc. Digital Audio Effects Workshop (DAFx*, 2003.

[14] S. Dubnov, "Generalization of spectral flatness measure for non-gaussian linear processes," *IEEE Signal Processing Letters*, vol. 11 n8, pp. 698–701, 2004.

[15] G. de Krom, "A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals," *Journal of Speech, Language, and Hearing Research*, vol. 36, no. 2, pp. 254–266, 1993.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[17] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio aignals," in *Proc. 14th Int. Conf. Digit. Signal Process. (DSP2002)*, vol. 2, (Santorini), pp. 967–970, July 2002.

[18] R. M. Gray and J. Berger, "Detection and modeling of transient audio signals with prior information," 2005.

[19] ITU, *Method for the subjective assessment of intermediate quality level of coding systems (ITU-T Recommendation ITU-R BS.1534-1)*. International Telecommunications Union, 2003.

[20] J. Liebetrau, S. Schneider, and T. Sporer, "Statistics of mushra revisited," in *Proceedings of the 127th Audio Engineering Society Convention (AES 127)*, no. 7825, Oct. 2009.

[21] E. Lesaffre, D. Rizopoulos, and R. Tsonaka, "The logistic transform for bounded outcome scores," *Biostat*, vol. 8, no. 1, pp. 72–85, 2002.