A Sub-Band Approach to Modification of Musical Transients

Markus Zaunschirm
Joshua D Reiss
Anssi Klapuri

Computer Music Journal, Volume 36, Number 2, Summer 2012, pp. 23-36 (Article)

Published by The MIT Press

A Sub-Band Approach to Modification of Musical Transients

Markus Zaunschirm,* Joshua D. Reiss,† and Anssi Klapuri†

*University of Music and Performing Arts Graz
Institute of Electronic Music and Acoustics
Inffeldgasse 10/3, 8010 Graz, Austria
markus.zaunschirm@student.tugraz.at
†School of Electronic Engineering
and Computer Science
Queen Mary University of London
327 Mile End Road, London E1 4NS, UK
{josh.reiss, anssi.klapuri}@eecs.qmul.ac.uk

Abstract: The transient modifier is a type of audio effect that changes the level of the transient parts in a musical signal while leaving the steady-state parts unchanged. This article presents a high-performance algorithm for transient detection and modification, one that is 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 of transients are performed in the frequency-domain using a sub-band approach. Detection is based on both phase and energy information using an adaptive threshold, and modification is carried out independently at each sub-band. The performance of the proposed sub-band approach was compared with other transient-modification algorithms using subjective listening tests. We show that the sub-band approach with adaptive threshold mostly outperforms other approaches.

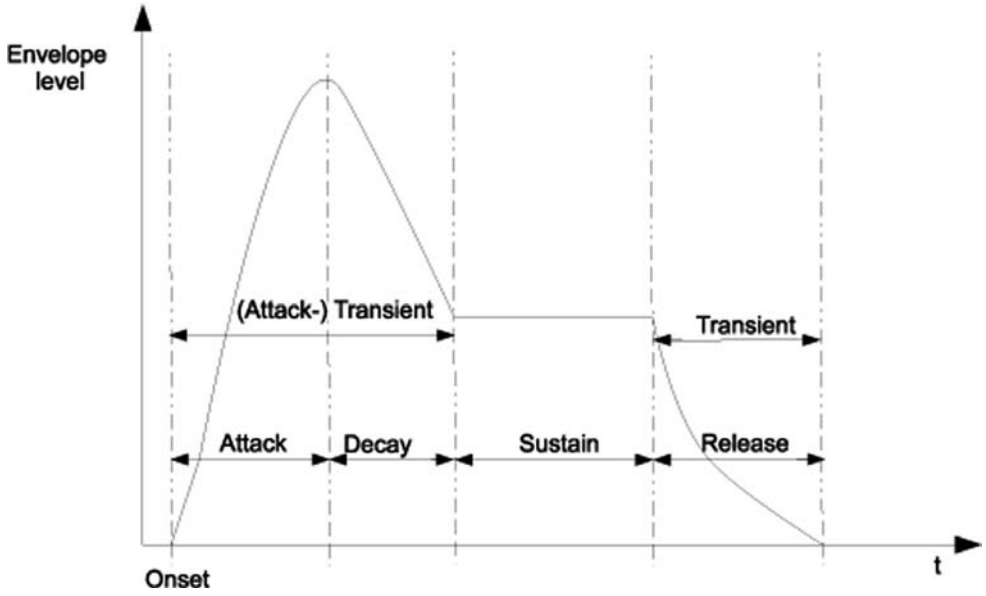
Musical transients are known for holding much of the perceptual information within musical tones. A change in the relative levels of the transient and steady-state parts of a musical tone significantly changes the perceived timbre of many instruments. Level changes of transient parts may be used to alter the dynamic range of a music piece. They can also change the perceptual attributes of the mix such as the "punchiness" or the perceived distance of the sources.

Unlike other dynamic processors such as compression or expansion, which react to the overall 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 artefacts. Two scenarios are worth special attention. The first is when the signal consists of soft transients, generated by non-percussive musical tones. Although 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 consists of overlapping notes generated by many sources. Finally, the transient modifier should be easy to use, with minimal manual intervention required, and capable of real-time implementation. These constraints provide the motivation for this work.

There is little published previous work on transient modification. Goodwin and Avendano (2006) presented two different approaches. The first algorithm uses a first-order energy difference of consecutive short-time Fourier transform (STFT) frames to detect transient segments, and then modifies the whole signal by applying a transient-dependent gain function to the original signal. The second algorithm uses the modulation spectrum of the audio signal, and is capable of altering transient components without explicit detection. Other methods of transient detection and modification have been based on sinusoidal modeling (Thornburg 2005) or on a sinusoidal-plus-transient-plus-noise model (Verma and Meng 2000). Further, transient detection is closely related to onset detection. A number of methods have been proposed for onset detection, including those based on the energy changes of the signal (Schloss 1985), the time derivative of the energy (spectral flux; Masri 1996; Duxbury, Sandler, and Davies 2002), the phase information (Bello and Sandler 2003), and the combination of both the energy and phase information (Duxbury et al. 2003).

Computer Music Journal, 36:2, pp. 23–36, Summer 2012

Figure 1. Idealized envelope evolution of a single note.



We propose a sub-band approach to transient modification. The advantage of this approach is that it allows estimating the degree of "transience" of different regions of the time/frequency plane, instead of merely transience as a function of time. As a result, modification can be targeted more accurately on the transient parts of the signal, while leaving the co-occurring steady sounds intact as far as possible.

This article is organized as follows. In the next section, we provide a definition of the transient in a musical audio signal. Then we describe our transient-detection function. The possible frequency-domain modifications are stated next. Based on the results and assumptions of the previous sections, we present an implementation of the entire audio effect and discuss its real-time aspects. To evaluate the quality of the suggested implementation, we perform several listening tests. The experiments and obtained results are summarized in the last section.

Definition of Transient

Informally, transients can be defined as short-time intervals during which the signal evolves quickly and unpredictably (Bello et al. 2005). 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 onset and attack parts are also closely related to the concept of a transient.

Figure 1 illustrates what is meant by a transient in this article. In sound synthesis, musical tones are often segmented into parts known as attack, decay, sustain, and release. *Transients* are the time intervals during which the signal characteristics change abruptly. As can be seen, this can take place during the attack, decay, and release parts. The onset of a sound is the time instant that marks the beginning of the temporally extended attack (transient), and is the earliest time at which the transient can be detected reliably. Note that the physical onset is often different from the perceived onset time. According to Vos and Rasch (1981) the physical onset marks the starting time of the stimulus, whereas the perceived onset (also known as perceptual attack) marks the time instance at which the stimulus is first perceived. This relative

difference is explained by the fact that the maximum level of a note is often reached just after a gradual level increase.

Complex music signals can be seen as a combination of steady-state and transient parts. This is emphasized by spectral modeling synthesis (Serra and Smith 1990), where deterministic (steady-state) parts are represented by sinusoids with slowly varying parameter trajectories and the stochastic (transient) parts are modelled by a filtered noise component. It is stated by Thornburg (2005) that these transient parts are commonly characterized by abrupt changes in amplitudes, phases, or frequencies, rapid decays in amplitudes, and "fast transitions" in both frequency and amplitude. Most of these changes occur when a new note is played, and thus are due to incoming energy associated with note onsets. As depicted in Figure 1, these attack transients are a combination of the attack and decay part.

Because one application of transient modification is to enhance the perceived impact on the sounding body (e.g., drum strike or piano hammer strike), it is crucial to detect transients associated with the onset. Other possible transient parts are rapid decays in amplitude, 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. Thus, negative energy changes also need to be detected. Fast transitions in both frequency and amplitude are mostly related to diverse means of expression (e.g., vibrato). For many applications, these expressive parts should remain unmodified in order to preserve the artistic content of the music signals. Thus, we want to identify transients associated with the note onset and the decay which can occur after the attack part, measure their duration, and apply an appropriate modification.

Transient Detection

Detection functions can be used to identify the transient parts. In general, a robust detection function will have large values (near 1.0) in transient regions, and small values (near 0.0) elsewhere. To arrive at an appropriate transient-detection function we analyzed methods for measuring transience that have been used in onset detection (Bello and Sandler 2003; Bello et al. 2005; Dixon 2006), spectral models (Serra and Smith 1990), speech modeling (Makhoul 1975), and speech enhancement (de Krom 1993; Dubnov 2004). We tried to find the most suitable method in terms of reliable identification of the transient parts, with low computational complexity, high-temporal precision, and using a suitable signal representation for high-quality transient modification. A full description of the tested approaches can be found in Zaunschirm (2010).

The complex-domain onset-detection function (Duxbury et al. 2003) was chosen because it fulfilled the requirements mentioned earlier. This function uses an STFT and retains both phase and amplitude information. The STFT of the input signal $x(n)$ is defined as

$$X(n,k) = \sum_{m=-N/2}^{N/2-1} x(nh+m)w(m)e^{-2i\pi mk/N}, \qquad (1)$$

where n and k are the time and frequency index, i is the imaginary unit, $w(m)$ is an N-point window, and h is the hop size between adjacent windows. In the detection function, each Fourier coefficient is given as a combination of its magnitude and phase $X_k(n) = |X_k(n)|\, e^{i\varphi_k(n)}$, where $|X_k(n)|$ is the magnitude and φ_k the phase of the k^{th} bin at time n. The predicted target Fourier coefficient is $\hat{X}_k(n) = |\hat{X}_k(n)|\, e^{i\hat{\varphi}_k(n)}$, where the target magnitude $|\hat{X}_k(n)|$ is the magnitude of the previous STFT frame $|X_k(n-1)|$, and the target phase $\hat{\varphi}_k(n)$ is calculated from the phase of the two previous STFT frames according to the phase vocoder principle (Fischman 1997):

$$\hat{\varphi}_k(n) = princ\arg\left[2\varphi_k(n-1) - \varphi_k(n-2)\right], \qquad (2)$$

where φ_k corresponds to the unwrapped phase of the k^{th} frequency bin and *princ*arg maps the values of the deviation in the range of $[-\pi, \pi]$. The latter is defined as a modulo operation where the divisor and remainder, per convention, share the same sign:

$$princ\arg(\Delta\varphi_k) = \mathrm{mod}(\Delta\varphi_k + \pi, -2\pi) + \pi. \qquad (3)$$

The transient-detection function is then given by the Euclidean distance between the predicted and actual-measured complex Fourier coefficient for each frequency bin k: A frame-by-frame detection function is defined by:

$$T_o(n) = \sum_{k=1}^{N} \sqrt{D_k(n)}. \quad (4)$$

With:

$$D_k(n) = \{[\Re(\hat{X}_k(n)) - \Re(X_k(n))]^2 + [\Im(\hat{X}_k(n)) - \Im(X_k(n))]^2\}, \quad (5)$$

where $\Re(\cdot)$ and $\Im(\cdot)$ denote the real and the imaginary parts, respectively. The transient values are bounded between [0,1] using a simple peak follower

$$T(n) = T_o(n)/\gamma(n), \quad (6)$$

with

$$\gamma(n) = \max_{j\in\{0,1,\ldots\ldots,n\}}(T_o(j)) \quad (7)$$

capable of real-time operation.

For locally steady-state regions, the frequency and amplitude should remain constant, $T(n) \cong 0$, and for highly transient regions, $T(n) \cong 1$. The combination of phase and magnitude information enables accurate detection of "pronounced" (percussive) and "non-pronounced" (pitched) transients for both multi-voiced and single instrument signals. Further confirmation of this approach is given in Bello et al. (2005) and Bello and Sandler (2003).

Transient Duration

A median-filtered detection function can be used to obtain an adaptive threshold for the detection of impulsive noise in music signals (Kauppinen 2002), which is comparable to transient detection. The adaptive threshold is defined as

$$\vartheta(n) = a \cdot \underset{j\in J_n}{median}(|T(j)|), \quad (8)$$

where a is a scaling factor and $J_n = \{n - \Delta n, n - \Delta n + 1, \ldots n + \Delta n\}$ In order to prevent the threshold from rising at the position of a peak, the length of the median filter, $2\Delta n + 1$, has to be set longer than the assumed duration of the peak in the transient-detection function. So the length of the median filter depends on the assumed maximum transient duration (140 msec throughout this article) and the time resolution $\Delta t = h/f_s$ of the detection function, where f_s is the sampling rate and h is the hop size (in samples) between successive windows. The estimated transient regions, along with the detection function, adaptive threshold, and onsets, are shown in Figure 2 for a sample of popular music.

Transient Modification

Consider a complex Fourier coefficient for the kth frequency bin at time instance n as a combination of its magnitude and phase:

$$X_k(n) = |X_k(n)|\, e^{j\varphi_k(n)}. \quad (9)$$

If the signal is considered transient, the magnitude of all bins should be modified accordingly. Let us define relative transience as the difference between the transient value $T(n)$ and the threshold value $\vartheta(n)$:

$$\tau(n) = T(n) - \vartheta(n). \quad (10)$$

The actual modification value $G(n)$ is a function of $\tau(n)$ and user input g:

$$G(n) = F(\tau(n), g). \quad (11)$$

High $\tau(n)$ indicates transient regions more reliably than a small $\tau(n)$. In order to minimize the effect of false positives, the modification should be dependent on the difference between the transient value and the threshold value. A smaller difference will result in less modification. A linear function was defined that maps the values of the relative transience between [0, 0.5] linearly onto modification values between $[1, g]$. For low relative transience($\tau(n) < 0.1$), however, the modification value should remain near 1 (no modification). Further, we observed that the relative transience rarely exceeds 0.3. Based on this, a soft mapping function was also defined. Figure 3 depicts the linear function (solid line) and soft

Figure 2. Estimated transient parts using complex-domain detection and an adaptive threshold with and without look-ahead, with $\Delta t = 10$ msec *and* $\Delta n = 14$.

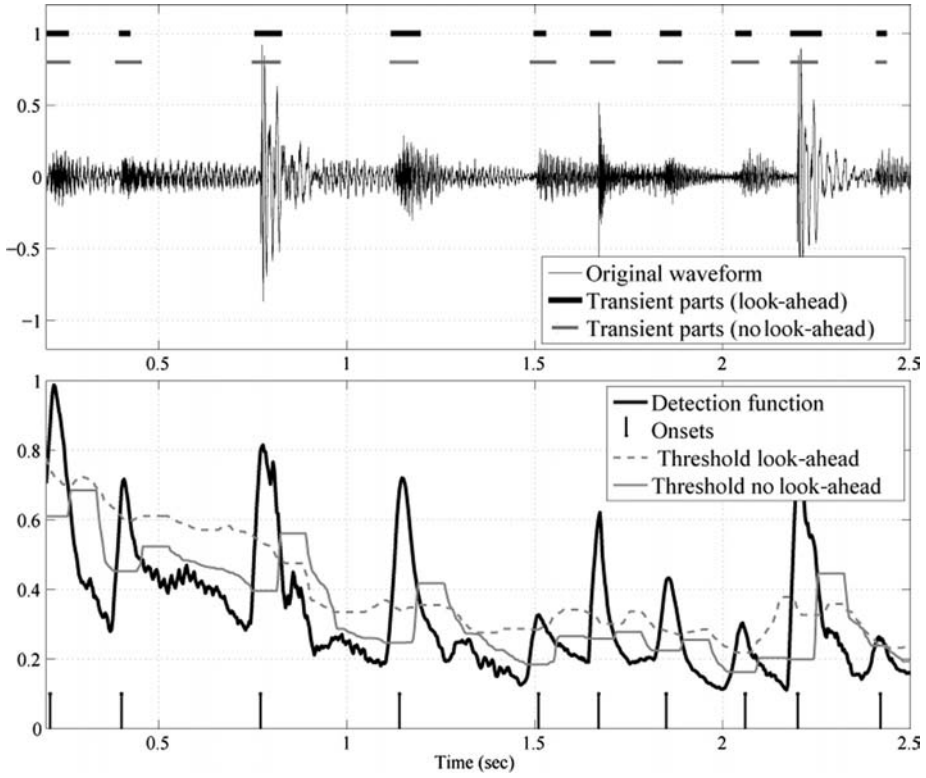


Figure 2

Figure 3. Modification values G as a function of relative transience τ *for a transient amplification scenario in which the maximal gain parameter* g *is set to 1.2.*

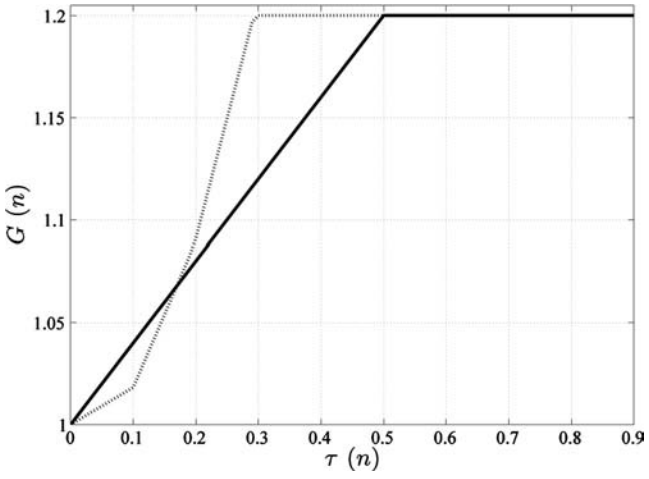


Figure 3

mapping function (dashed line) for the amplification scenario and $g = 1.2$.

The frequency-domain transient-based modification was introduced in Goodwin and Avendano (2006). Thereby, the modified complex Fourier coefficient $\tilde{X}_k(n)$ is obtained by a function of the actual modification value $G(n)$ and the original Fourier coefficient $X_k(n)$. Goodwin and Avendano distinguished between two modification schemes: linear modification, where the magnitude of the original coefficient is multiplied by the corresponding modification gain value:

$$\left|\tilde{X}_k(n)\right| = \left|X_k(n)\right| G(n), \tag{12}$$

and nonlinear modification, which uses a slightly different computation:

$$|\tilde{X}_k(n)| = (|X_k(n)| + 1)^{G(n)^2} - 1 \tag{13}$$

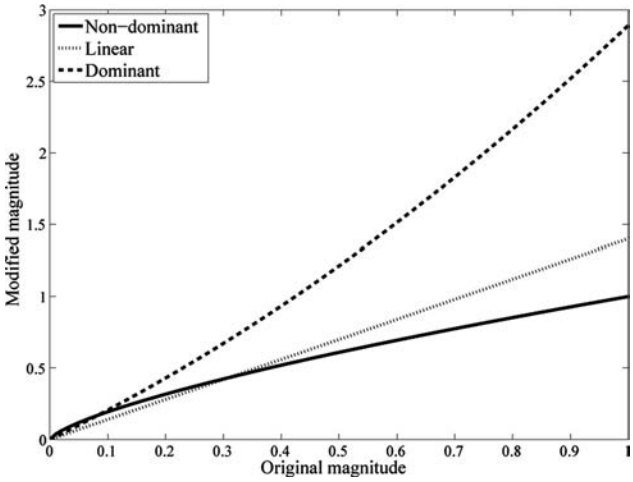


Figure 4. Modifications of output magnitude for a fixed modification value G = 1.4. *Note that "Linear" corresponds to Equation 12, "Dominant" to Equation 13, and "Non-dominant" to Equation 14.*

Goodwin and Avendano stated that the nonlinear modification yields modifications that sound more natural, and stated that dominant spectral components are more affected by the nonlinearity because, in a complex mix, transient parts are assumed to be dominant over stable parts. According to Rodet and Jaillet (2001), transients are more noticeable at high frequencies, and these frequencies usually carry less energy than low frequencies, so we also defined another nonlinear modification scheme that more strongly affects spectral components that have less energy:

$$\left|\tilde{X}_k(n)\right| = \left|X_k(n)\right|^{\frac{1}{G(n)^{1/3}}}. \qquad (14)$$

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 Figure 4.

Sub-Band Processing

If we consider polyphonic or multi-voiced music, one note may be in its transient part at the same time point that another note is in its stable part. Accordingly, if the transient part at this time point is detected, a modification of the whole spectrum would include an inappropriate modification of the overlaid stable part. Further, when considering a single note, the attack times for different frequencies (harmonics) are assumed to have different durations. Klapuri (1999) states that especially low-frequency 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.

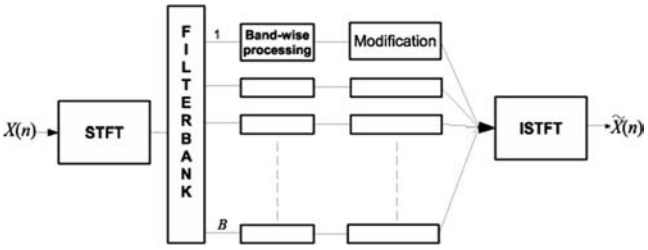


Figure 5. Overview of the sub-band transient detection and modification system.

To overcome these problems, the frequency-domain transient detection and modification was implemented as a sub-band approach. Each sub-band in each window is characterized as being steady state or transient. The modification can be performed flexibly, because every sub-band has its own transient function for detection and weighting. The general scheme of the system is shown in Figure 5. The different stages of the detection and modification process are discussed next.

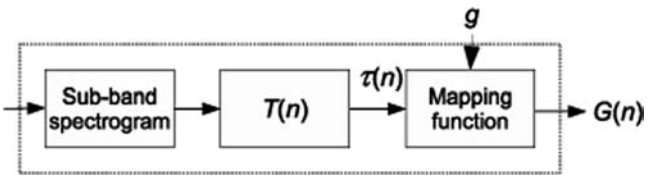
Audio Input

The input signal can be a monaural or stereo signal at any sampling rate. In order to compare results with the original file, the maximum level of the input signal is reduced to leave headroom for the modification. In the case of stereo files, transient detection and modification is done separately for each channel.

STFT

Time resolution has a crucial influence on the detection and modification. We know that transient portions generally have short duration, assumed to be between 30 and 130 msec. On the other hand, the frequency resolution should also be high to get

Figure 6. Processing stage.



better results. As a compromise, we used an STFT frame size of $N = 2{,}048$ and a hop size of $L = 512$ samples (a time resolution of 11.6 msec at $f_s = 44{,}100$ Hz).

Filter Bank

The resulting spectrogram can be split into several non-overlapping sub-bands according to the logarithmic scale between the frequency range of 20 Hz to 20 kHz. We used $B = 6$ sub-bands, leading to a bandwidth of roughly 1.5 octaves per band. The sub-band processing is depicted in Figure 6.

Summarized, transient-detection functions $T(n)$ are generated according to Equations 4 and 5 and adapted to the human auditory system by weighting them with the total energy (over all sub-bands) at each time instance. This is based on the fact that intensity differences are perceived relative to the overall intensity. Adaptive thresholds for each sub-band are obtained from Equation 8, with an empirically derived scaling factor $a = 1.2$ and filter length $2\Delta n + 1 = 29$. The computed relative transience values $\tau(n)$ are mapped according to the soft tuning characteristics to obtain the transient modification values $G(n)$. Figure 7 shows the transient modification values for a suppression scenario and $g = 0.6$. The actual modification can be performed as stated in Equations 12, 13, and 14.

User Settings

In order to allow the user to adjust the behavior 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. The factor $\mathrm{mod}_{\min}[dB]$ defines the lowest amount of modification that is realized for $G(n) \neq 1$. So if a modification would result in a level change lower than $\mathrm{mod}_{\min}[dB]$, the modification is not performed. A high setting of this parameter may be used if only hard transients are to be modified. The parameters a and Δn can be used to change the behavior of the median filter; a changes the overall level of the threshold and Δi the filter length. For higher a, only strong transients will be detected. Figure 8 shows the original, modified, and residual signal (defined as original signal minus modified signal) for a suppression scenario. Demonstrations of the modified samples using all modification schemes (see the Transient Modification section) are available (Zaunschirm 2010).

Real-Time Aspects

In general, it is possible to implement the detection and modification in real time. The introduced latency is mainly determined by the STFT, the needed overlap, the corresponding inverse Fourier transform, and the design of the median filter for threshold generation. According to Equation 8, median filtering requires a look-ahead time of Δn frames. To reduce overall latency, a median filtering approach without look-ahead can be used. This scheme also affects the detection and modification behavior of the entire audio effect, however. The possible impact on the detected transients and their durations are depicted in Figure 2. It can be seen that the two threshold schemes are comparable and also detect the same transient parts, but with slightly different durations.

Listening Test and Results

A listening test was carried out to compare four different configurations of the described method to each other and to a consumer VST plug-in (Sonnox Ltd. 2007). The tested configurations were the sub-band approach with an adaptive threshold, a single-band approach with an adaptive threshold,

Figure 7. Relative transience and resulting modification values (according to Equation 13) for a pop music sample.

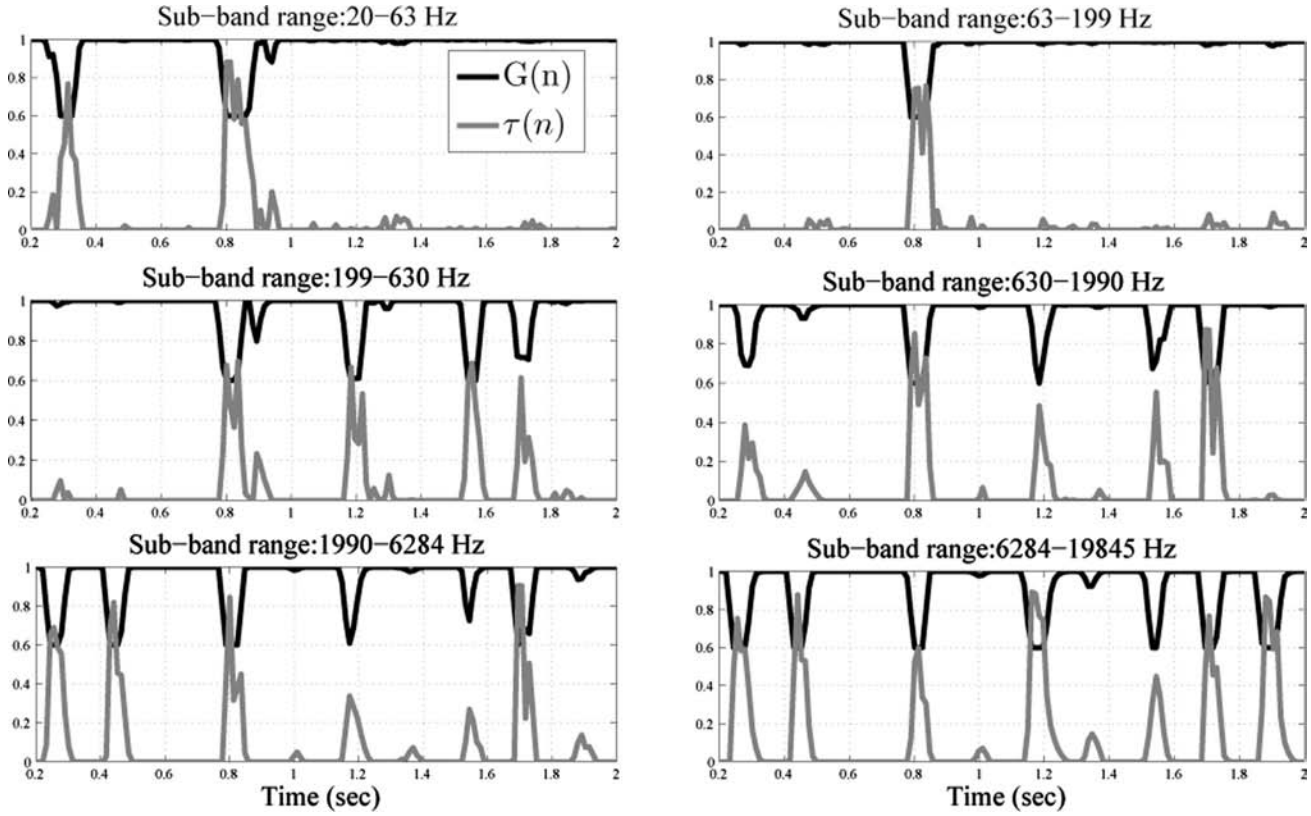
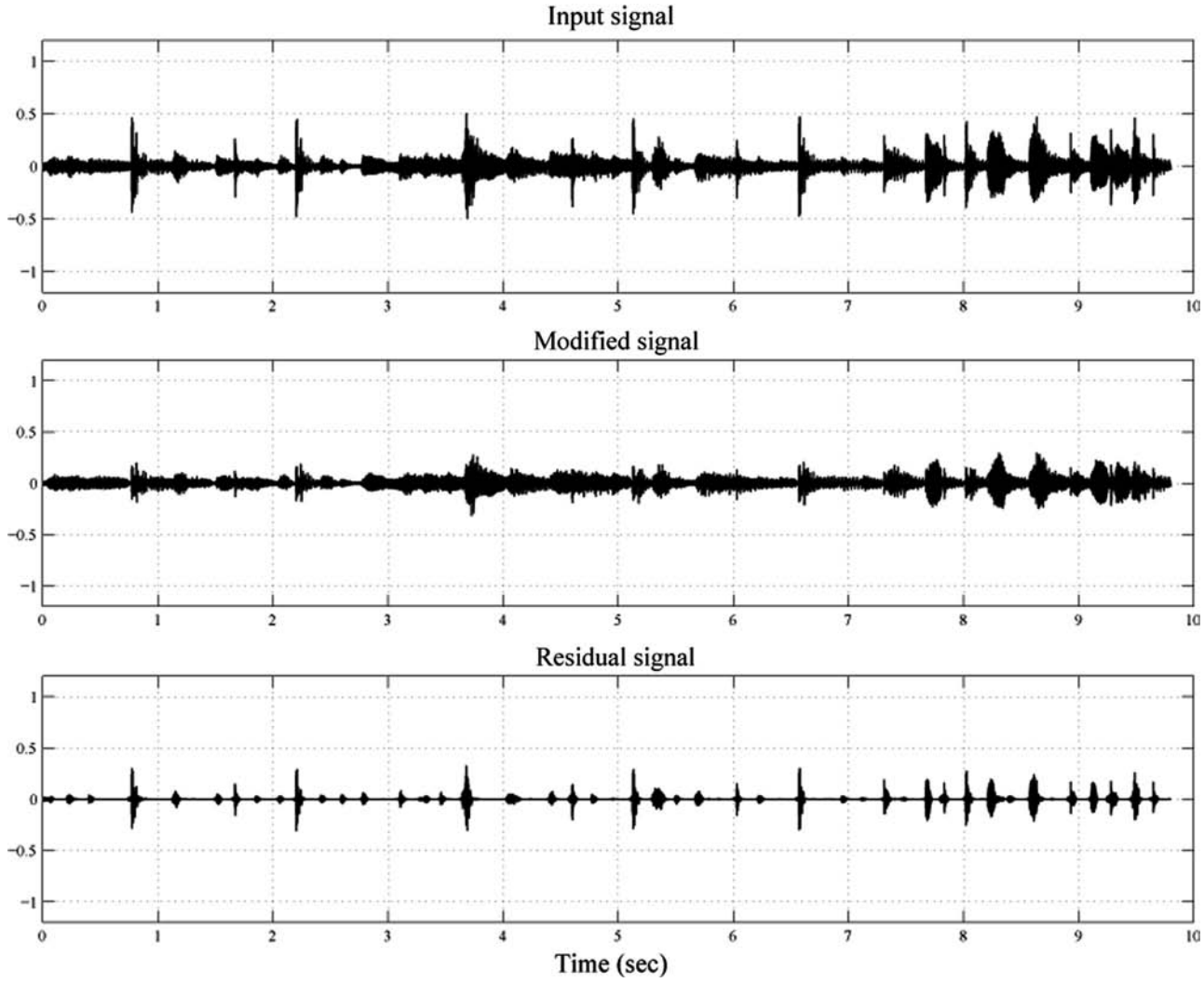


Table 1. Global Parameters Used for Transient Modification

Parameter	*Suppression*	*Amplification*
g	0.4–0.99	1.01–1.8
$\text{mod}_{\text{min}}[dB]$	0–4	0–6
a	1–1.5	1–1.5
Filter length $2\Delta n + 1$	10–40	10–40

an all-band approach with adaptive threshold (every frequency bin is treated as a different sub-band), and a single-band approach with fixed threshold (our implementation of Goodwin and Avendano [2006]). The reference signals were chosen from different music genres. In order to get a diversified sample pool, we used percussive, pitched-percussive, and non-percussive sounds, in monophonic as well as polyphonic polytimbral contexts. All samples were modified using each of the different approaches. The parameters were set as constant as possible to ensure a fair, non-discriminatory comparison. The listening test assessed the audibility of the implemented effect and evaluated the quality of the resulting audio output for the different approaches. We also tested the change of the perceptual attribute "punchiness" or "forcefulness" and the perception of distance for different amounts of modification. The listeners were asked to rate the samples according to (1) the perceived transient suppression, (2) the ability to modify transients from all types of sound sources, (3) the punchiness, (4) the perceived distance, (5) the ability to amplify the transients while not affecting the steady-state portions, and (6) the modification quality. The full instructions and samples are presented in Zaunschirm (2010).

Figure 8. Audio input, resulting modified output, and residual signal for transient suppression scenario (according to Equation 13) of a complex mix; $g = 0.6$.

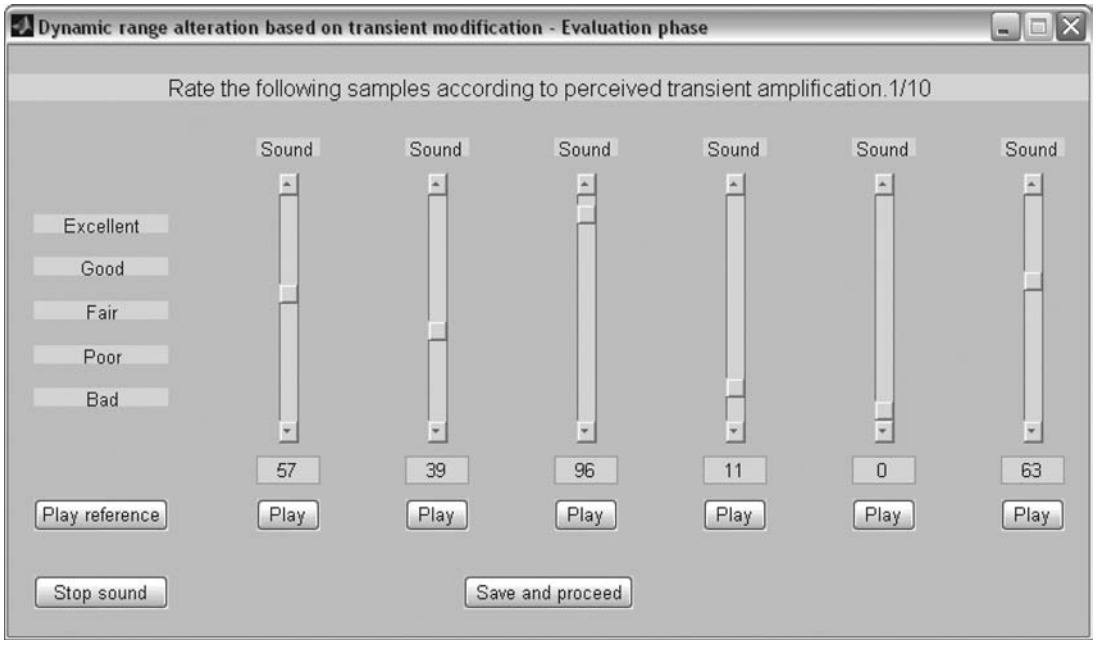


Method

Because transient modification intentionally affects 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 to achieve a similar maximum modification for all approaches under test. The consumer VST plug-in offers a variety of different parameter settings: we set the input gain to zero, the overshoot value to near the middle of the scale (about 8 msec), and the amount of modification (ratio) to a low value 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 (ITU-R BS.1534-1;

Figure 9. Interface of the listening test.



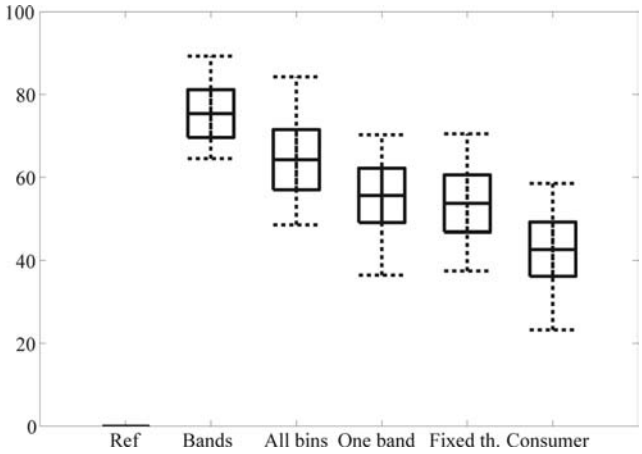
ITU 2003). Participants in MUSHRA tests are presented with sets of processed audio excerpts and asked to rate their basic audio quality compared to an unprocessed reference audio excerpt. Usually, each excerpt set includes the unprocessed audio as hidden reference and a 3.5 kHz low-pass-filtered version of the excerpt as a low-quality anchor. As this was designed for the subjective assessment of intermediate audio quality and not for audio effect evaluation, the low-pass-filtered original was not used as an anchor in our tests. For evaluation of the ability to amplify or suppress the transients, the hidden reference was used as anchor. But this anchor was not appropriate for evaluation of the modification quality. We used anchors showing similar types of impairments as the output of the single-band, fixed-threshold approach. So in general, the test may be considered as a mixture of a MUSHRA-test and a semantic differential or rating, because we did not use both the hidden reference and the anchor in all cases.

The test included eight experiments. The order of the experiments and of the excerpts within each trial was randomized. Each experiment contained six signals to be graded, from 3 to 15 seconds long. The subjects could listen to the signals in any order, any number of times. The grading scale was continuous, from bad (grade 0) to excellent (grade 100). The interface is shown in Figure 9. (The instructions for each experiment were given on an additional sheet.)

Participants

We recruited 13 experienced music listeners (9 men and 4 women, ages 24–34 years). Tests were performed using headphones and took 30–40 minutes to complete, including the initial training phase and the actual evaluation phase. For the pre-screening, we made sure that all tested participants had normal hearing. For post-screening, we performed a combination of numerical test and manual inspection. We calculated the Pearson's r and Spearman's φ for each participant, which correlated 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. Further, sets in which the participants did not rate the hidden reference consistently were also considered to be outliers (for rating the quality, the hidden reference

Figure 10. Evaluation of ability to modify transients from all sources, showing mean and 95-percent confidence interval according to a t distribution and whiskers extending to 25th and 75th percentiles. Ref = ratings for the hidden reference; Bands = the sub-band approach using six sub-bands; All bins = the all-bins approach, where each bin is treated as a different sub-band; One band = the one-band approach with adaptive threshold; Fixed th. = the one band approach with fixed threshold; Consumer = the samples generated using a consumer VST plug-in.



should be rated 100; for rating the perceived suppression/amplification and changed "punch," the hidden reference should be rated 0). Because participants tend to treat MUSHRA-related tasks as a ranking task and therefore slightly penalize the hidden reference if they misidentify the signal with the highest quality (Liebetrau, Schneider, and Sporer 2009; in our test, this means that they misidentify other signals as the signal with no 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.

Results

With the exception of evaluation of perceived distance, each figure in this section shows, from left to right, the ratings for the hidden reference (Ref), the sub-band approach using six sub-bands (Bands), the all-bins approach, where each bin is treated as a 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 plug-in (Consumer).

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, because results are assumed to be more meaningful in this representation (Lesaffre, Rizopoulos, and Tsonaka 2002).

Table 2. Identified Outliers for Each Experiment

Experiment	*Outliers*
modification of all sources	1
transient suppression	1
increased punch	1
not affecting steady state	0
perceived quality	3

Ability to Modify Transient Parts from all Types of 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. As a test sample we used a complex pop music sample. The results are shown in Figure 10.

According to the mean value of the ratings, the sub-band approach performs best for the modification of transients from all sources. The results of the *t*-test for this experiment imply that there is no significant difference between the results, except between the results of the sub-band and consumer ratings.

Figure 11. Evaluation of perceived transient suppression, showing mean and 95-percent confidence interval according to a t *distribution and whiskers extending to the 25th and 75th percentiles. Ref = ratings for the hidden reference; Bands = the sub-band approach using six sub-bands; All bins = the all-bins approach, where each bin is treated as a different sub-band; One band = the one-band approach with adaptive threshold; Fixed th. = the one band approach with fixed threshold; Consumer = the samples generated using a consumer VST plug-in.*

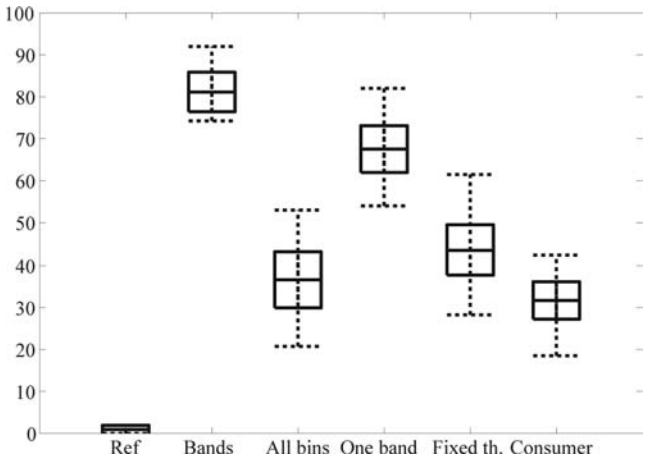
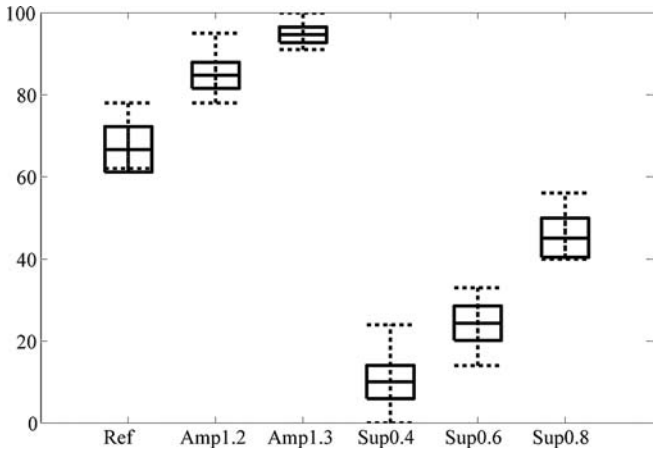


Figure 12. Evaluation of perceived distance, showing mean and 95-percent confidence interval according to a t distribution and whiskers extending to the 25th and 75th percentiles; with indicated amplification or suppression amount. Ref = no modification; Amp1.2 = amplification with g = 1.2; Amp1.3 = amplification with g = 1.3; Sup0.4 = suppression with g = 0.4; Sup0.6 = suppression with g = 0.6; Sup0.8 = suppression with g = 0.8.



Perceived Transient Suppression

For this experiment we used a slap bass sample as hidden reference signal. The participants were asked to rate the approaches according to the perceived transient suppression—e.g., if the bass seems to be played softer. We can see in Figure 11 that the bands and one-band approach perform significantly better than the other approaches.

Increased Punch

The hidden 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 from 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 being very near, and 0 being far away. We used a conga sample as reference. 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 Figure 12, it is possible to change the perceived microphone-source distance by changing the relation between transient and steady-state parts (e.g., a greater amplification of transients is perceived as a closer distance).

Modification of Transients while not Affecting Steady-State Portions

As hidden reference signals, we used (1) a drum sample with added stable sinusoids and (2) a bowed string. Participants were asked to rate to what extent the steady-state portions of the signal were modified. This experiment was intended to spot which approach is best able to detect transient parts and their durations. It should also verify the ability of the bands approach to not affect steady-state portions in the presence of a transient part when they are located in different sub-bands. The sub-band approach and the consumer tool equally outperform the other three approaches. The difference between the two best performing approaches was not significant, however; see Figure 13.

Figure 13. Evaluation of ability to modify transients while not affecting steady-state portions, showing mean and 95-percent confidence interval according to a t distribution and whiskers extending to the 25th and 75th percentiles. Ref = ratings for the hidden reference; Bands = the sub-band approach using six sub-bands; All bins = the all-bins approach, where each bin is treated as a different sub-band; One band = the one-band approach with adaptive threshold; Fixed th.= the one band approach with fixed threshold; Consumer = the samples generated using a consumer VST plug-in.

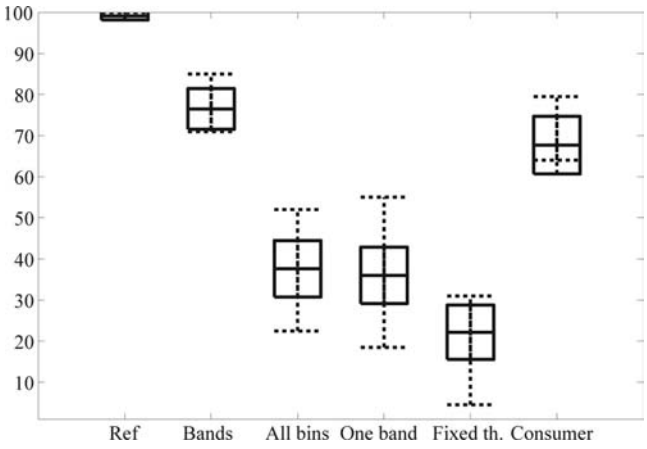
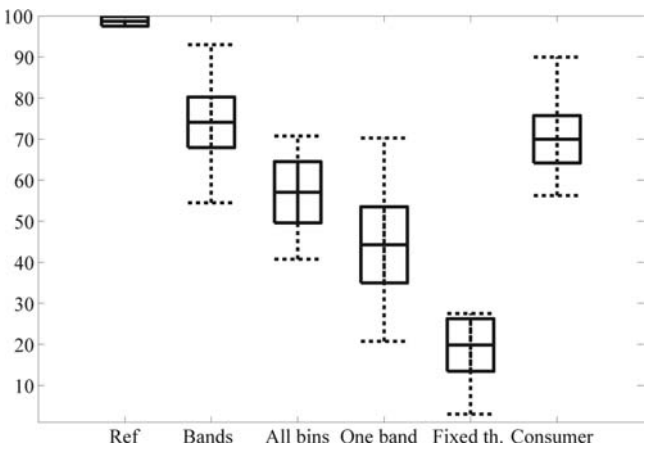


Figure 14. Evaluation of perceived quality, showing mean and 95-percent confidence interval according to a t distribution and whiskers extending to the 25th and 75th percentiles. Ref = ratings for the hidden reference; Bands = the sub-band approach using six sub-bands; All bins = the all-bins approach, where each bin is treated as a different sub-band; One band = the one-band approach with adaptive threshold; Fixed th. = the one band approach with fixed threshold; Consumer = the samples generated using a consumer VST plug-in.



Perceived Quality

As hidden references, we used a mix of three instruments (drums, piano, and bass) and a simpler mix of two instruments (guitar and 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-band approach and consumer tool are significantly rated as having the least impairment of quality. We can also see in Figure 14 that the fixed-threshold approach introduces the most unwanted effects.

Conclusion

In this article a new, high-performance transient modifier was developed and evaluated. The chosen approach to transient detection used a complex-domain onset-detection function with sub-bands and a short-time Fourier transform in order to modify only the bands and time intervals that have 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 were discussed that were implemented with either a look-ahead (having increased latency) or median filtering (having inaccuracies in the measurement of transient duration).

MUSHRA-style listening tests were performed in order to compare this approach against other approaches for several performance measures. We showed that the sub-band, adaptive-threshold approach outperformed other 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 might be possible to improve this performance using a multi-resolution STFT or a constant-Q transform.

Although real-time implementations were discussed and analyzed, 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. For example, it would be possible to apply different gain factors to different sub-bands (similar to filtering or shelving), or to apply diverse audio effects to both the transient or steady-state parts separately.

References

Bello, J., and M. Sandler. 2003. "Phase-Based Note Onset Detection for Music Signals." In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. v–441.

Bello, J., et al. 2005. "A Tutorial on Onset Detection in Music Signals." *IEEE Transactions on Speech and Audio Processing* 13(5):1035–1047.

de Krom, G. 1993. "A Cepstrum-Based Technique for Determining a Harmonics-to-Noise Ratio in Speech Signals." *Journal of Speech, Language, and Hearing Research* 36(2):254–266.

Dixon, S. 2006. "Onset Detection Revisited." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 133–137.

Dubnov, S. 2004. "Generalization of Spectral Flatness Measure for Non-Gaussian Linear Processes." *IEEE Signal Processing Letters* 11(8):698–701.

Duxbury, C., M. Sandler, and M. Davies. 2002. "A Hybrid Approach to Musical Note Onset Detection." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 33–38.

Duxbury, C., et al. 2003. "Complex Domain Onset Detection for Musical Signals." In *Proceedings of the International Conference on Digital Audio Effects*, pp. 1–4.

Fischman, R. 1997. "The Phase Vocoder: Theory and Practice." *Organised Sound* 2(2):127–145.

Goodwin, M. M., and C. Avendano. 2006. "Frequency-Domain Algorithms for Audio Signal Enhancement Based on Transient Modification." *Journal of the Audio Engineering Society* 54(9): 827–840.

ITU. 2003. Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems (ITU-T Recommendation ITU-R BS.1534-1). Geneva: International Telecommunications Union.

Kauppinen, I. 2002. "Methods for Detecting Impulsive Noise in Speech and Audio Signals." In *Proceedings of the 14th International Conference on Digital Signal Processing*, vol. 2, pp. 967–970.

Klapuri, A. 1999. "Sound Onset Detection by Applying Psychoacoustic Knowledge." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 115–118.

Lesaffre, E., D. Rizopoulos, and R. Tsonaka. 2002. "The Logistic Transform for Bounded Outcome Scores." *Biostat* 8(1):72–85.

Liebetrau, J., S. Schneider, and T. Sporer. 2009. "Statistics of Mushra Revisited." In *Proceedings of the 127th Audio Engineering Society Convention*, Paper 7825 (pages unnumbered).

Makhoul, J. 1975. "Linear Prediction: A Tutorial Review." *Proceedings of the IEEE* 63(4):561–580.

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

Rodet, X., and F. Jaillet. 2001. "Detection and Modeling of Fast Attack Transients." In *Proceedings of the International Computer Music Conference*, pp. 30–33.

Schloss, W. A. 1985. *On the Automatic Transcription of Percussive Music: From Acoustic Signal to High Level Analysis*. PhD thesis, Stanford University.

Serra, X., and J. Smith. 1990. "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition." *Computer Music Journal* 14(4):12–24.

Sonnox Ltd. 2007. Sonnox Oxford Transient Modulator. Available online at www.sonnoxplugins.com/pub/plugins/products/transmod.htm. Accessed 4 December 2011.

Thornburg, H. 2005. *Detection and Modeling of Transient Audio Signals with Prior Information*. PhD thesis, Stanford University.

Verma, T. S., and T. H. Y. Meng. 2000. "Extending Spectral Modeling Synthesis with Transient Modeling Synthesis." *Computer Music Journal* 24(2):47–59.

Vos, J., and R. Rasch. 1981. "The Perceptual Onset of Musical Tones." *Attention, Perception, and Psychophysics* 29(4):323–335.

Zaunschirm, M. 2010. "Transient Modification Report." Queen Mary University of London. Available online at www.elec.qmul.ac.uk/digitalmusic/audioengineering/transientmodification. Accessed 4 December 2011.