Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/TR 10-58

Perception-based clipping of audio signals[1]

Bruno Defraene[2] [3], Toon van Waterschoot[2], Hans Joachim Ferreau[2], Moritz Diehl[2] and Marc Moonen[2]

August 2010
Published in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO 2010)*, Aalborg, Denmark, Aug. 2010, pp. 517-521

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.be* in the directory *pub/sista/bdefraen/reports/10-58.pdf*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SCD(SISTA), Kasteelpark Arenberg 10, 3001 Leuven, Belgium, Tel. +32 16 321788, Fax +32 16 321970, WWW: *http://homes.esat.kuleuven.be/∼bdefraen*. E-mail: *bruno.defraene@esat.kuleuven.be*.

[3]This research work was carried out at the ESAT Laboratory of Katholieke Universiteit Leuven, in the frame of K.U.Leuven Research Council CoE EF/05/006 (“Optimization in Engineering (OPTEC)”), and the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, “Dynamical systems, control and optimization”, 2007-2011). The scientific responsibility is assumed by its authors.

PERCEPTION-BASED CLIPPING OF AUDIO SIGNALS

Bruno Defraene, Toon van Waterschoot, Hans Joachim Ferreau, Moritz Diehl and Marc Moonen

Dept. E.E./ESAT, SCD-SISTA, Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
phone: +32 16 321788, fax: +32 16 321970
email: bruno.defraene@esat.kuleuven.be

ABSTRACT

Clipping is an essential signal processing operation in real-time audio applications. Still, existing clipping techniques introduce a considerable amount of distortion which results in a significant degradation of perceptual sound quality. In this paper, we propose a novel approach to clipping which aims to minimize perceptible clipping-induced distortion. The clipping problem is formulated as a sequence of constrained optimization problems, all of which can be solved numerically in a very efficient way. A comparative evaluation of the presented "perception-based" clipping technique and existing clipping techniques is performed using two objective measures of perceptual sound quality. For both measures, the application of the perception-based clipping technique results in consistently higher scores as compared to existing clipping techniques.

1. INTRODUCTION

In many audio devices, the amplitude of a digital audio signal cannot exceed a maximum level. This amplitude level restriction often necessitates a *clipping* operation to be performed on the audio signal. Clipping consists of attenuating incoming signal sample amplitudes such that no sample amplitude exceeds the maximum level (referred to as *clipping level* from here on). However, such a clipping operation may introduce different kinds of unwanted distortion : harmonic distortion, intermodulation distortion, and aliasing distortion [1]. These additional frequency components, which were not present in the original frequency spectrum, then reduce the perceptual quality of the audio signal.

Most existing clipping techniques make use of a static nonlinearity acting on the input audio signal in a sample-by-sample fashion. These clipping techniques are thus governed by a fixed input-output characteristic, mapping a range of input amplitudes to a reduced range of output amplitudes. Depending on the sharpness of the input-output characteristic, one can distinguish two types of clipping techniques. A first type is *hard clipping*, where the input-output characteristic exhibits an abrupt ("hard") transition from the linear zone to the nonlinear zone. In a series of listening experiments performed on normal hearing subjects [2] and hearing-impaired subjects [3], it is concluded that the application of hard clipping to audio signals has a large negative effect on perceptual sound quality scores, irrespective of the subject's hearing acuity. A second type of clipping techniques is *soft clipping*, where the input-output characteristic exhibits a gradual ("soft") transition from the linear zone to the nonlinear zone. The actual shape of the input-output characteristic can vary, and different soft clipping input-output characteristics have been proposed (e.g. see [4]). In the above cited listening experiments, it is concluded that the application of soft clipping to audio signals has a smaller negative effect on perceptual sound quality scores, again irrespective of the subject's hearing acuity.

The outlined traditional clipping techniques are basically inflexible in that each input signal sample is processed independently using a fixed input-output characteristic. In this paper, in contrast, we propose a more flexible approach to clipping, enabling to adapt to the instantaneous properties of the input signal. Our perception-based clipping approach builds upon recent advances in the fields of psychoacoustics and numerical optimization. First, incorporating knowledge of human perception of sounds (psychoacoustics) appears indispensable for achieving minimal perceptible clipping-induced distortion. In other applications of audio processing, this has proven to be successful, e.g. in perceptual audio coding [5] and audio signal requantization [6]. Secondly, the clipping problem is formulated as a sequence of constrained optimization problems, which necessitate efficient numerical solution algorithms.

The paper is organized as follows. In Section 2, clipping is formulated as a sequence of constrained optimization problems. Section 3 deals with efficiently solving these optimization problems. In Section 4, results of a comparative evaluation of the presented perception-based clipping technique and existing clipping techniques are discussed. Finally, Section 5 presents concluding remarks.

2. OPTIMIZATION PROBLEM FORMULATION

Figure 1 schematically depicts the operation of the perception-based clipping technique presented here. A digital input audio signal $x[n]$ is segmented into frames of N samples, with an overlap length of P samples between successive frames. Processing of one frame x_k consists of the following steps :

1. Calculate the instantaneous global masking threshold $t_k \in \mathbb{R}^{\frac{N}{2}+1}$ of the input frame x_k
2. Calculate output frame $y_k^* \in \mathbb{R}^N$ as the solution of an op-

This research work was carried out at the ESAT laboratory of Katholieke Universiteit Leuven, in the frame of K.U.Leuven Research Council CoE EF/05/006 Optimization in Engineering (OPTEC) and the Belgian Programme on Interuniversity Attraction Poles initiated by the Belgian Federal Science Policy Office IUAP P6/04 (DYSCO, 'Dynamical systems, control and optimization', 2007-2011) and supported by the Research Foundation-Flanders (FWO).

© EURASIP, 2010 ISSN 2076-1465

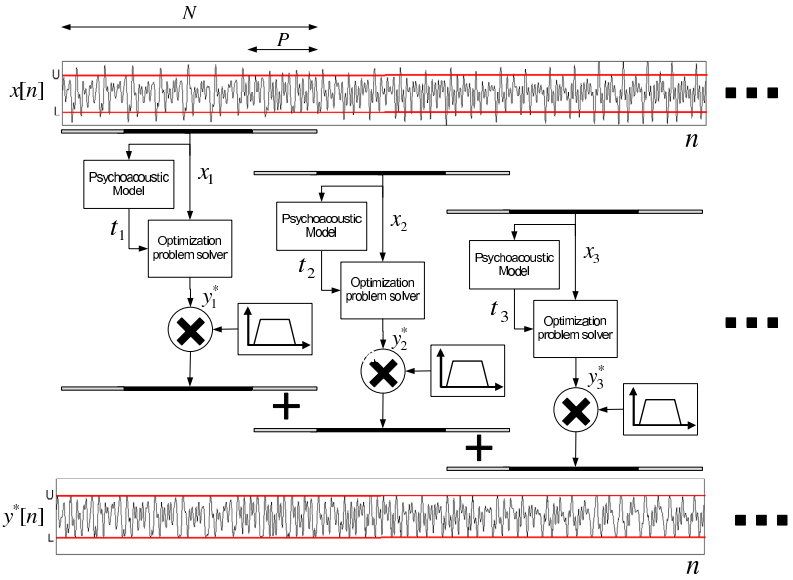


Figure 1: Schematic overview of the perception-based clipping technique

timization problem

3. Apply trapezoidal window to output frame y_k^* and sum output frames to form a continuous output audio signal $y^*[n]$.

In the next subsections, the different processing steps will be discussed in more detail.

2.1 Convex quadratic program

The core of the perception-based clipping technique consists in calculating the solution of a constrained optimization problem for each frame. From the knowledge of the input frame x_k and its instantaneous properties, the output frame y_k^* is calculated. Let us define the optimization variable of the problem as y_k, the output frame. A necessary constraint on the output frame y_k is that the amplitude of the output samples cannot exceed the upper and lower clipping levels U and L. The cost function we want to minimize must reflect the amount of perceptible distortion added between y_k and x_k. We can thus fomulate the optimization problem as an inequality constrained frequency domain weighted L2-distance minimization, i.e.

$$y_k^* = \underset{y_k \in \mathbb{R}^N}{\arg\min} \frac{1}{2} \sum_{i=0}^{N-1} w_k(i) |Y_k(e^{j\omega_i}) - X_k(e^{j\omega_i})|^2 \quad \text{s.t.} \quad l \leq y_k \leq u \tag{1}$$

where $\omega_i = (2\pi i)/N$ represents the discrete frequency variable, $X_k(e^{j\omega_i})$ and $Y_k(e^{j\omega_i})$ are the discrete frequency components of x_k and y_k respectively, the vectors $u = U1_N$ and $l = L1_N$ contain the upper and lower clipping levels respectively (with $1_N \in \mathbb{R}^N$ a vector of all ones), and $w_k(i)$ are the weights of a perceptual weighting function to be defined in subsection 2.2. Notice that in case the input frame x_k does not violate the inequality constraints, the optimization problem (1) has a trivial solution $y_k^* = x_k$ and the input frame is transmitted unaltered by the clipping algorithm.

Formulation (1) of the optimization problem can be written as a standard quadratic program (QP) as follows [1]

$$\begin{aligned} y_k^* &= \underset{y_k \in \mathbb{R}^N}{\arg\min} \, (y_k - x_k)^H \, D^H W_k D \, (y_k - x_k) \quad \text{s.t.} \quad l \leq y_k \leq u \\ &= \underset{y_k \in \mathbb{R}^N}{\arg\min} \, \frac{1}{2} \, y_k^H \underbrace{D^H W_k D}_{\text{Hessian } H_k} \, y_k + \big(\underbrace{-D^H W_k D \, x_k}_{\text{Gradient } g = -H_k x_k} \big)^H \, y_k \\ &\qquad \text{s.t.} \quad l \leq y_k \leq u \end{aligned} \tag{2}$$

where $D \in \mathbb{C}^{N \times N}$ is the DFT-matrix defined as

$$D = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & e^{-j\omega_1} & e^{-j\omega_2} & \dots & e^{-j\omega_{N-1}} \\ 1 & e^{-j\omega_2} & e^{-j\omega_4} & \dots & e^{-j\omega_{2(N-1)}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & e^{-j\omega_{N-1}} & e^{-j\omega_{2(N-1)}} & \dots & e^{-j\omega_{(N-1)(N-1)}} \end{bmatrix} \tag{3}$$

and $W_k \in \mathbb{R}^{N \times N}$ is a diagonal weighting matrix with positive weights $w_k(i)$, obeing symmetry $w_k(i) = w_k(N-i)$ for $i = 1, 2, ..., \frac{N}{2} - 1$,

$$W_k = \begin{bmatrix} w_k(0) & 0 & 0 & \dots & 0 \\ 0 & w_k(1) & 0 & \dots & 0 \\ 0 & 0 & w_k(2) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & w_k(N-1) \end{bmatrix} \tag{4}$$

It can be shown that by imposing these requirements on the weighting matrix, the Hessian matrix H_k in (2) is guaranteed to be real and positive definite. Hence, formulation (2) defines a strictly convex QP. Many efficient solution algorithms have been presented to solve such QPs in a fast and reliable way, e.g. [7]. In Section 3, we willl show that by exploiting the structure of the problem, the QPs can be solved even more efficiently.

[1] The superscript H denotes the Hermitian transpose

2.2 Perceptual weighting function

In order for the cost function in (1) to represent the amount of perceptible distortion added between input frame x_k and output frame y_k, the perceptual weighting function w_k must be chosen judiciously. Distortion at certain frequency bins is more perceptible than distortion at other frequency bins. Two phenomena of human auditory perception are responsible for this,

- The *absolute threshold of hearing* is the required intensity (dB) of a pure tone such that an average listener will just hear the tone in a noiseless environment. The absolute threshold of hearing is a function of the tone frequency and has been measured experimentally [8].
- *Simultaneous masking* is a phenomenon of human auditory perception where the presence of certain spectral energy (the masker) masks the simultaneous presence of weaker spectral energy (the maskee).

Combining both phenomena, the instantaneous global masking threshold of a signal gives the amount of distortion energy (dB) at each frequency bin that can be masked by the signal. In this framework, consider the input frame x_k acting as the masker, and $y_k - x_k$ as the maskee. By selecting the weight $w_k(i)$ for the distortion term $|Y_k(e^{j\omega_i}) - X_k(e^{j\omega_i})|^2$ in the cost function (1) to be inversely proportional to the value of the global masking threshold of x_k at frequency bin i, the cost function reflects the amount of perceptible distortion introduced. This can be specified as

$$w_k(i) = \begin{cases} 10^{-\alpha t_k(i)} & \text{if } 0 \le i \le \frac{N}{2} \\ 10^{-\alpha t_k(N-i)} & \text{if } \frac{N}{2} < i \le N-1 \end{cases} \tag{5}$$

where t_k is the global masking threshold (in dB). Appropriate values for the compression parameter α are determined to lie in the range 0.04-0.06.

Part of the ISO/IEC 11172-3 MPEG-1 Layer 1 psychoacoustic model 1 [9] is used to calculate the instantaneous global masking threshold t_k of the input frame. A detailed description of the operation of this psychoacoustic model can be found in [5]. We will only outline the major steps in the computation of the instantaneous global masking threshold here :

1. *Identification of noise and tonal maskers*

   After performing a spectral analysis of the input frame x_k, tonal maskers and noise maskers are identified in the spectrum. The distinction between these two types of maskers is important as they have a different masking power.

2. *Calculation of individual masking thresholds*

   Each tonal masker and each noise masker has an individual masking effect on neighboring frequency regions. This masking effect can be represented by an individual masking threshold per masker.

3. *Calculation of global masking threshold*

   The input signal x_k consists of several tonal maskers and noise maskers. In this model, additivity of masking effects is assumed. Under this assumption, the instantaneous global masking threshold t_k can be calculated as the sum of the individual masking thresholds and the absolute threshold of hearing.

2.3 Trapezoidal window

To ensure continuity of the output audio signal $y^*[n]$, a trapezoidal window is applied to the output frame y_k^* before summation. Hence, in the overlap zone between two consecutive output frames, the output frames are crossfaded : the previous output frame fades out while the current output frame fades in. In this fashion, audible artefacts due to a lack of continuity in the output signal are greatly reduced.

3. OPTIMIZATION PROBLEM SOLUTION

An instance of the quadratic optimization problem (2) is solved numerically at each time step. Real-time operation of the clipping algorithm imposes very strict restrictions on the maximum problem solution time. For example, considering a frame length of N=512 samples and an overlap length of P=128 samples at a sampling frequency of 44.1 kHz, the time step is 8.7 ms. This means that a 512-dimensional QP is to be solved in every 8.7 ms. Since general-purpose QP solvers have shown to be inadequate to achieve sufficiently low solution times, real-time operation calls for an application-tailored solution strategy. A first step is to formulate the dual optimization problem of (2) as follows. First, the Lagrangian $\mathscr{L}(y_k, \lambda_{k,u}, \lambda_{k,l})$ of the QP is given by

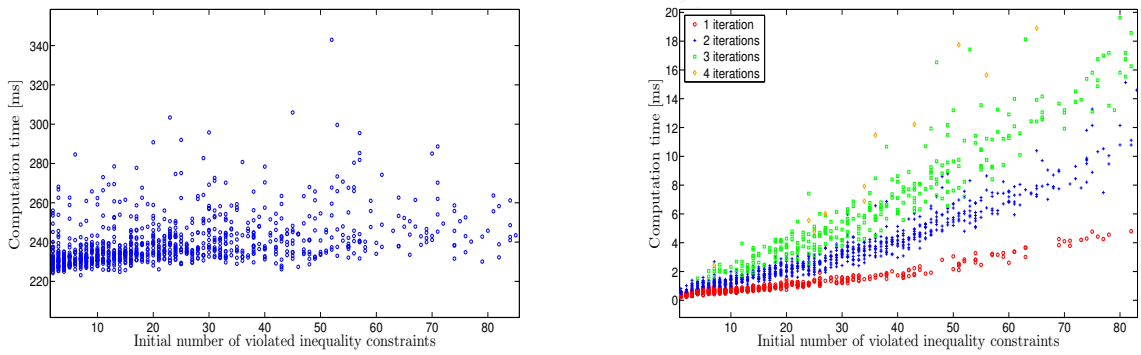
$$\begin{aligned} \mathscr{L}(y_k, \lambda_{k,u}, \lambda_{k,l}) &= \frac{1}{2}\,(y_k - x_k)^T\, H_k\,(y_k - x_k) + \lambda_{k,u}^T(y_k - u) \\ &\qquad + \lambda_{k,l}^T(l - y_k) \\ &= \frac{1}{2}\, y_k^T H_k y_k + (\lambda_{k,u} - \lambda_{k,l} - H_k x_k)^T y_k \\ &\qquad - \lambda_{k,u}^T u + \lambda_{k,l}^T l + \frac{1}{2}\, x_k^T H_k x_k \end{aligned} \tag{6}$$

where $\lambda_{k,u}, \lambda_{k,l} \in \mathbb{R}^N$ denote the vectors of Lagrange multipliers associated to the upper clipping level constraints and the lower clipping level constraints respectively. Then, the Lagrange dual function equals

$$\begin{aligned} q(\lambda_{k,u}, \lambda_{k,l}) &= \inf_{y_k} \mathscr{L}(y_k, \lambda_{k,u}, \lambda_{k,l}) \\ &= -\frac{1}{2}(\lambda_{k,u} - \lambda_{k,l} - H_k x_k)^T H_k^{-1}(\lambda_{k,u} - \lambda_{k,l} - H_k x_k) \\ &\qquad - \lambda_{k,u}^T u + \lambda_{k,l}^T l + \frac{1}{2}\, x_k^T H_k x_k \end{aligned} \tag{7}$$

where the last equality follows from the positive definiteness of H_k. Finally, the dual optimization problem can be formulated as

$$\begin{aligned} \lambda_k^* &= \arg\max_{\lambda_k}\ q(\lambda_k) \quad \text{s.t. } \lambda_k \ge 0 \\ &= \arg\max_{\lambda_k}\ -\frac{1}{2}(B\lambda_k - H_k x_k)^T H_k^{-1}(B\lambda_k - H_k x_k) - e^T C\lambda_k \\ &\qquad + \frac{1}{2}\, x_k^T H_k x_k \quad \text{s.t. } \lambda_k \ge 0 \\ &= \arg\min_{\lambda_k}\ \frac{1}{2}\lambda_k^T \underbrace{B^T H_k^{-1} B}_{\text{Hessian } \tilde{H}_k}\lambda_k + \underbrace{(C^T e - B^T x_k)}_{\text{Gradient } \tilde{g}}{}^T \lambda_k \quad \text{s.t. } \lambda_k \ge 0 \end{aligned} \tag{8}$$



(a) Full-scale QP

(b) Working set strategy

Figure 2: Scatter plot of optimization problem solution computation time vs. initial number of violated inequality constraints [GenuineIntel CPU 2826 Mhz, using qpOASES [10]]

where $\lambda_k \in \mathbb{R}^{2N}$, $B \in \mathbb{R}^{N \times 2N}$ and $C \in \mathbb{R}^{N \times 2N}$ are defined as

$$\lambda_k = \begin{bmatrix} \lambda_{k,u} \\ \lambda_{k,l} \end{bmatrix} \tag{9}$$

$$B = \begin{bmatrix} I_n \mid -I_n \end{bmatrix} \tag{10}$$

$$C = \begin{bmatrix} UI_n \mid -LI_n \end{bmatrix} \tag{11}$$

Computation of y_k^* is straightforward,

$$\begin{aligned} y_k^* &= -H_k^{-1}(B\lambda_k^* - H_k x_k) \\ &= x_k - H_k^{-1} B\lambda_k^* \end{aligned} \tag{12}$$

Optimization problem (2) can be solved efficiently by exploiting the fact that only few of the large number ($2N$) of inequality constraints are expected to be active in the solution (see [11] for a similar idea). A two-level external active set strategy is adopted, where the following steps are executed in each outer iteration :

1. Check which inequality constraints are violated in the previous solution iterate. In case no inequality constraints are violated, the algorithm terminates.
2. Add these violated constraints to an active set S of constraints to watch.
3. Solve a small-scale QP corresponding to (8) with those $\lambda_k(i)$ not in S set to zero. Evaluation of eq. (12) yields the new solution iterate.

Using this strategy, the solution of optimization problem (2) is found by solving several small-scale QPs instead of by solving the full-scale QP at once. Simulations show that more than 4 iterations are rarely necessary. In Figure 2, solution computation times for the proposed working set strategy are compared to the scenario of solving the full-scale QP. For both solution strategies, solution computation times of many instances of QP (2) (with N=512 variables) are plotted against the initial number of violated inequality constraints. In Figure 2(a), solution computation times for the full-scale QP can be seen to lie in the range 220-350 ms. In Figure 2(b), solution computation times for the working set strategy can be seen to increase with increasing number of constraint violations and with increasing number of necessary iterations. A reduction of computation time with a factor ranging from 10 up to 200 is achieved. Moreover, the real-time restriction of 8.7 ms is met for the majority of the QP instances solved.

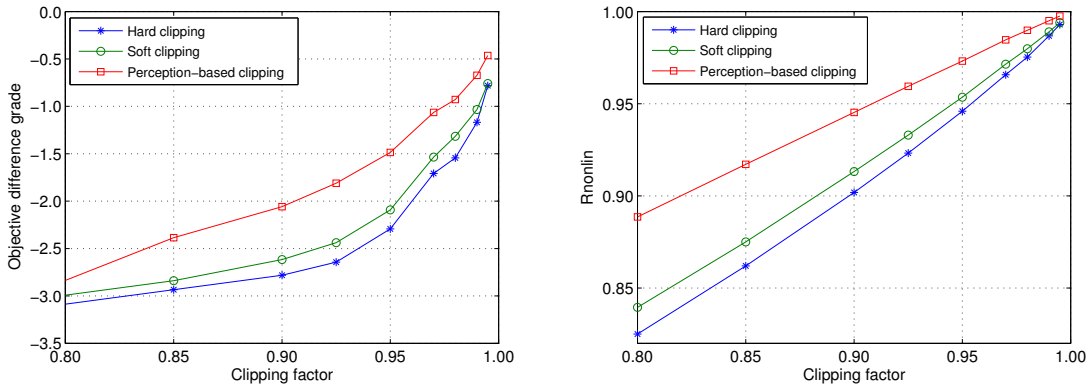
4. EVALUATION

For sound quality evaluation purposes, eight audio signals (16 bit mono @44.1 kHz) of different musical styles and with different maximum amplitude levels were collected. Each signal was processed by three different clipping techniques :

- Hard symmetrical clipping (with $L = -U$)
- Soft symmetrical clipping as defined in [4]
- Perception-based clipping, with parameter values N=512, P=256, $\alpha = 0.06$

This was performed for nine clipping factors {0.80, 0.85, 0.90, 0.925, 0.950, 0.97, 0.98, 0.99, 0.995}, where the clipping factor is defined as 1-(fraction of signal samples exceeding the upper or lower clipping level). From the clipping factor, a corresponding clipping level U can be derived for each signal.

For each of a total of 216 processed signals, two objective measures of sound quality are calculated. An objective measure of sound quality predicts the subjective quality score attributed by an average human listener. A first objective measure of sound quality is calculated using the Basic Version of the PEAQ (Perceptual Evaluation of Audio Quality) standard [12, 13]. Taking the reference signal and the signal under test as an input, PEAQ calculates an objective difference grade on a scale of 0 (imperceptible impairment) to -4 (very annoying impairment). One should note that PEAQ was designed in particular for predicting the performance of audio codecs, and that PEAQ quality scores are reported to correlate less well with subjective quality scores for some other applications (e.g. [14]). Therefore, also a second objective measure of sound quality is calculated. Rnonlin is a perceptually relevant measure of nonlinear distortion, for which correlations as high as 0.98 between objective and subjective ratings have been obtained [15]. Rnonlin decreases with increasing perceptible distortion (1 = no perceptible distortion).



(a) PEAQ Basic Version

(b) Rnonlin

Figure 3: Average objective sound quality scores vs. clipping factor for hard clipping, soft clipping and perception-based clipping

The results of this comparative evaluation are shown in Figure 3. In Figure 3(a), the obtained average PEAQ objective difference grade over eight audio signals is plotted as a function of the clipping factor, and this for the three different clipping techniques. Analogously, Figure 3(b) shows the results for the Rnonlin measure. The obtained results for both measures are seen to be in accordance with each other. As expected, we observe a monotonically increasing average sound quality score for increasing clipping factors. Soft clipping is seen to result in slightly higher objective sound quality scores than hard clipping for all considered clipping factors. Clearly, the perception-based clipping technique is seen to result in significantly higher objective sound quality scores than the other clipping techniques.

5. CONCLUSION

In this paper, we have developed a novel approach to clipping. Clipping of an audio signal was formulated as a sequence of constrained optimization problems aimed at minimizing perceptible clipping-induced distortion. A comparative evaluation of the presented perception-based clipping technique and existing clipping techniques was performed using two objective measures of perceptual sound quality. For both measures, the application of the presented clipping technique was observed to result in consistently higher scores as compared to existing clipping techniques.

REFERENCES

[1] F. Foti, "Aliasing distortion in digital dynamics processing, the cause, effect, and method for measuring it: The story of 'digital grunge!'," in *Preprints AES 106th Conv.*, Munich, Germany, May 1999, Preprint no. 4971.

[2] C.-T. Tan, B. C. J. Moore, and N. Zacharov, "The effect of nonlinear distortion on the perceived quality of music and speech signals," *J. Audio Eng. Soc.*, vol. 51, no. 11, pp. 1012–1031, Nov. 2003.

[3] C.-T. Tan and B. C. J. Moore, "Perception of nonlinear distortion by hearing-impaired people," *Int. J. Audiol.*, vol. 47, pp. 246–256, May 2008.

[4] A. N. Birkett and R. A. Goubran, "Nonlinear loudspeaker compensation for hands free acoustic echocancellation," *Electron. Lett.*, vol. 32, no. 12, pp. 1063–1064, Jun. 1996.

[5] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proc. IEEE*, vol. 88, no. 4, pp. 451–515, Apr. 2000.

[6] D. De Koning and W. Verhelst, "On psychoacoustic noise shaping for audio requantization," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc.*, Hong Kong, Apr. 2003, pp. 413–416.

[7] H. J.Ferreau, H. G. Bock, and M. Diehl, "An online active set strategy to overcome the limitations of explicit MPC," *Int. J. Robust Nonlinear Contr.*, Jul. 2008.

[8] E. Terhardt, "Calculating virtual pitch," *Hearing Res.*, vol. 1, no. 2, pp. 155 – 182, 1979.

[9] ISO/IEC, "11172-3 Information technology - Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s - Part 3: Audio," 1993.

[10] H. J. Ferreau, "qpOASES software package," http://www.qpoases.org, 2007–2010.

[11] E. Polak, H. Chung, and S. S. Sastry, "An external active-set strategy for solving optimal control problems," University of California, Berkeley, Tech. Rep. EECS-2007-90, Jul. 2007.

[12] International Telecommunications Union Recommendation BS.1387, "Method for objective measurements of perceived audio quality," 1998.

[13] T. Thiede et al., "PEAQ: The ITU standard for objective measurement of perceived audio quality," *J. Audio Eng. Soc.*, vol. 48, no. 1–2, pp. 3–29, Feb. 2000.

[14] A. de Lima et al., "Reverberation assessment in audioband speech signals for telepresence systems," in *Int. Conf. Signal Process. Multimedia Applic.*, Porto, Portugal, Jul. 2008, pp. 257–262.

[15] C.-T. Tan, B. C. J. Moore, N. Zacharov, and V.-V. Mattila, "Predicting the perceived quality of nonlinearly distorted music and speech signals," *J. Audio Eng. Soc.*, vol. 52, no. 7–8, pp. 699–711, Jul. 2004.