INCREASING LOUDNESS IN AUDIO SIGNALS: A PERCEPTUALLY MOTIVATED APPROACH TO PRESERVE AUDIO QUALITY

A. Jeannerot[1,2], N. de Koeijer[1,3], P. Martínez-Nuevo[1], M. B. Møller[1], J. Dyreby[1], and P. Prandoni[2]

[1] Bang & Olufsen, Acoustics R&D, Struer, Denmark
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] Delft University of Technology (TU Delft), Delft, Netherlands

ABSTRACT

We present a method to maintain the subjective perception of volume of audio signals and, at the same time, reduce their absolute peak value. We focus on achieving this without compromising the perceived audio quality. This is specially useful, for example, to maximize the perceived reproduction level of loudspeakers where simply amplifying the signal amplitude, and hence their peak value, is limited due to already constrained physical designs. In particular, we minimize the absolute peak value subject to a constraint based on auditory masking. This limits the perceptual difference between the original and the modified signals. Moreover, this constraint can be tuned and allows to control the resulting audio quality. We show results comparing loudness and audio quality as a function of peak reduction. These results suggest that our method presents the best trade-off between loudness and audio quality when compared against classical methods based on compression and clipping.

Index Terms— Loudness increase, crest factor reduction, peak reduction, perceptual models, audio signal processing

1. INTRODUCTION

It is usually desirable for loudspeakers, specially the ones small in size, to be able to increase loudness as much as possible while avoiding any distortion. This may be achieved by simply amplifying the amplitude of the audio signals. However, this is usually limited by physical and electronic design constraints. For example, the maximum displacement of the voice coil of a loudspeaker driver so that it still operates in a linear regime, or the maximum input level of a voltage amplifier so that clipping is avoided. Instead, it is common to resort to processing techniques that lead to an increase of the perceived volume level even in the presence of these limitations. For ease of explanation, we focus on the increase of loudness of voltage signals in this paper. The same principles apply in order to constrain the loudspeaker driver displacement. In this case, a linear model could be readily incorporated [1], [2].

Indirect methods to increase loudness operate by reducing the amplitude excursions of the audio signal and then, if ap-plicable, reamplifying to increase its energy. There exist two main classical approaches in the literature in this regard: dynamic range compression (DRC) [3] and clipping techniques [4]. The DRC reduces the amplification gain of the input signal for input levels above a predefined threshold. This reduction is achieved through a compression ratio. Additionally, it incorporates integration periods, referred to as attack and release times, thus introducing memory in the system. Even with a careful choice of these parameters, a DRC may inadvertently lead to poor audio quality [5].

Another common approach is the use of clipping. This is essentially a memoryless input-output map with a linear characteristic around the origin and nonlinear regions for larges input values. Depending on the characteristics of the nonlinear regime, it is possible to further divide clipping into hard and soft clipping. Hard clipping is characterized by clamping the output to a constant value when the input exceeds a predefined threshold. Soft clipping makes this transition smoother. Similar to DRCs, clipping typically results in audio quality degradation [5]. In [6], an approach is presented to reduce the impact of clipping by using a simplified masking model. They focus on minimizing the perceptual distance, based on noise masking, between the input and output signals such that the values of the latter lie within a predefined range.

In this paper, we present a perceptually motivated approach that increases the loudness of an audio signal while controlling the degradation in audio quality. As opposed to previous methods, even those incorporating perceptual features, we focus on minimizing the peak value of the input signal under a constraint based on the perceptual detectability of the resulting distortion. This constraint, based on multitonal masking, includes a parameter that is perceptually interpretable, referred to as *detectability*. This provides us with a way of tuning and controlling the output audio quality explicitly. This is particularly relevant in an environment that needs guarantees on the perceived audio quality of the output signals.

The paper is structured as follows: in section 2, we place our proposed perceptually-optimized loudness increase method and classical clipping techniques within an optimiza-

tion framework. Section 3 introduces the perceptual model used. Finally, in section 4, we demonstrate the performance of the proposed method by using loudness, audio quality, and crest factor reduction as evaluation criteria. We use a dataset of music clips consisting of kick drums and compare against DRCs and clipping techniques.

2. LOUDNESS INCREASE: AN OPTIMIZATION PERSPECTIVE

Instead of directly optimizing for loudness, it is usually more tractable to resort to indirect methods. The perception of volume is not only driven by the peak value of sound pressure waves but, to a large extent, it is also determined by their energy. Thus, in a scenario where maximum amplitude levels are limited, it may be possible to increase loudness by both minimizing the peak value and maximizing the energy. Later, it is always possible to reamplify the signal to its maximum amplitude level, thus increasing loudness. This can be formalized in terms of the so-called crest factor of a signal $\mathbf{x}$, i.e., $\mathrm{CR} := 10 \log_{10}(||\mathbf{x}||_\infty/||\mathbf{x}||_2)$ where $\mathbf{x} \in \mathbb{R}^N$ for some $N > 0$.

Unfortunately, minimizing the crest factor entails solving a nonconvex problem. In order to have stronger theoretical guarantees, we can settle instead for a suboptimal problem that is actually convex. We can achieve this by minimizing the ℓ_∞-norm subject to be close to the original signal in an energy sense. Thus, we minimize the peak value and aim at preserving the energy. We show in this section how this approach leads to the classical notions of hard clipping. Finally, our proposed method naturally follows: the energy constraint is weighted by a perceptual criterion instead. This constraint is known as detectability and provides us with control over the resulting audio quality.

2.1. Non-perceptual Optimization Approach

Minimizing the crest factor involves minimizing the ℓ_∞-norm of a signal and, at the same time, maximizing the ℓ_2-norm. However, it is more tractable to settle for suboptimal methods. Instead, it is possible to constrain error, in the ℓ_2 sense, and minimize the peak of a signal. In particular, we can write

$$S_\lambda(\mathbf{x}_0) := \underset{\mathbf{x} \in \mathbb{R}^N}{\arg\min} \quad ||\mathbf{x}||_\infty + \frac{1}{2\lambda}||\mathbf{x} - \mathbf{x}_0||_2 \quad (1)$$

for a signal $\mathbf{x}_0 \in \mathbb{R}^N$ and some $\lambda > 0$. Interestingly, this corresponds to performing hard clipping on $\mathbf{x}_0$. The connection comes from the fact that $S_\lambda(\mathbf{x}_0)$ is the proximal operator of $\lambda|| \cdot ||_\infty$.

In order to show this, notice that hard clipping can be computed from the soft-thresholding operator T_λ [7], i.e., $S_\lambda(\mathbf{x}_0) = \mathbf{x}_0 - T_\lambda(\mathbf{x}_0)$ where $T_\lambda(x) := ||x| - \lambda| \cdot \mathrm{sgn}(x)$ for $|x| > \lambda$ and $T_\lambda(x) := 0$ otherwise. Note that $T_\lambda(\mathbf{x}) \equiv$

$(T_\lambda(x_i))_{i=1}^N$. It follows then that $S_\lambda(x) = x$ for $|x| < \lambda$ and $S_\lambda(x) = \lambda$ otherwise. Similarly, $S_\lambda(\mathbf{x}) \equiv (S_\lambda(x_i))_{i=1}^N$.

As shown later, the main issue with this approach is that it leads to poor audio quality. However, we show in the next section how it is possible to include a frequency-domain weighting in the energy difference in such a way that this degradation in audio quality is controlled.

2.2. Perceptually Motivated Approach

Instead of simply focusing on solving (1), we resort to frequency masking. In particular, we introduce the matrix $\mathbf{P}_{\mathbf{x}_0}$ to compute the perceptual difference, or error, between the original and modified signals [8]. This error is constrained by a parameter $c \geq 0$, referred to as detectability, that provides information about how noticeable the difference is. This results in the following optimization problem

$$\begin{aligned} \min \quad & ||\mathbf{x}||_\infty \\ \mathrm{s.t.} \quad & ||\mathbf{P}_{\mathbf{x}_0}\mathbf{W}(\mathbf{x} - \mathbf{x}_0)||_2 \leq c \end{aligned} \quad (2)$$

where $\mathbf{W} \in \mathbb{R}^{N \times N}$ is the Discrete Fourier Transform (DFT) matrix. We explain in detail in the next sections how $\mathbf{P}_{\mathbf{x}_0}$ is computed and what the detectability parameter c implies. This form of the optimization problem is particularly interesting since the detectability provides us with a direct way of controlling the audio quality, reducing the peak value, and potentially increasing loudness. We also explore in this paper an alternative formulation, inspired by [6], i.e.,

$$\begin{aligned} \min \quad & ||\mathbf{P}_{\mathbf{x}_0}\mathbf{W}(\mathbf{x} - \mathbf{x}_0)||_2 \\ \mathrm{s.t.} \quad & ||\mathbf{x}||_\infty \leq \lambda \end{aligned} \quad (3)$$

where $\lambda > 0$ is the maximum amplitude level allowable in the system.

3. PSYCHOACOUSTIC MODEL

We use a model of human auditory perception that exploits tonal masking. In particular, the distortion detectability $D(\mathbf{x}_0, \mathbf{e})$ predicts how noticeable a sinusoidal distortion $\mathbf{e} \in \mathbb{R}^N$ is in the presence of a masking signal $\mathbf{x}_0 \in \mathbb{R}^N$ [8]. This distortion detectability can be expressed in a compact form as an ℓ_2-norm in the frequency domain [9]:

$$D(\mathbf{x}_0, \mathbf{e}) := ||\mathbf{P}_{\mathbf{x}_0}\hat{\mathbf{e}}||_2^2 \quad (4)$$

where $\hat{\mathbf{e}} := \mathbf{W}\mathbf{e}$ and $\mathbf{P}_{\mathbf{x}_0} \in \mathbb{R}^{N \times N}$ is a diagonal matrix that weighs the frequency components of $\hat{\mathbf{e}}$ based on the psychoacoustic masking properties of $\hat{\mathbf{x}}_0$. Thus, increasing values of D indicate a more noticeable distortion. This matrix $\mathbf{P}_{\mathbf{x}_0}$ is constructed in the following manner [9]:

$$\mathbf{P}_{\mathbf{x}_0} := \mathrm{diag}\left(\sqrt{\sum_{i=1}^{} \frac{c_s\hat{\mathbf{h}}_i^2}{||\hat{\mathbf{h}}_i\hat{\mathbf{x}}_0||_2^2 + c_a}}\right) \quad (5)$$

where $\hat{\mathbf{h}}_i \in \mathbb{R}^N$ represents the ith filter in a Gammatone filter bank modeling the outer- and middle-ear transfer functions as well as the i-th auditory filter (note that $\cdot^2$ operates element-wise). The constants c_a and c_s are calibration coefficients chosen such that $D(\mathbf{x}_0, \mathbf{e}) = 1$, i.e., when $\mathbf{e}$ is just detectable in the presence of $\mathbf{x}_0$ [8].

In our setting, we constrain the value of $D(\mathbf{x}_0, \mathbf{x} - \mathbf{x}_0)$. In particular, we want to find a signal $\mathbf{x}$ with a lower peak value in magnitude such that the detectability of the error signal $\mathbf{x} - \mathbf{x}_0$ is limited. We interpret the detectability of this error as the perceptual difference between the original and modified signals. As shown in the next section, although the resulting signal may have lower energy as well as maximum amplitude value, its crest factor is reduced and loudness is increased if we add a simple reamplification factor.

4. SIMULATIONS AND RESULTS

We focus on enhancing signals mainly composed of low-frequency content. This is of particular interest in reproduction systems since the human hearing system is less sensitive to this frequency range. Our analysis focuses on kick drum sounds sampled from a drum kit, which are characterized by a low-pass spike-like waveform followed by a resonant part. Clearly, naively increasing their loudness by linear amplification can lead to saturation of the voltage amplifier. Since they are often repeated several times across a song, focusing on single kick drum clips is a first step toward increasing the loudness of longer samples. In our experiments, we apply the method proposed in both formulations, as in (2) and (3), to 16 clips taken from [10] with a maximum duration of 1 s and downsampled to 1 kHz. We assess the performance using loudness, audio quality, and crest factor reduction as function of peak reduction as the underlying physical problem is the reduction of the loudspeaker excursion. The dataset used to generate the plots can be accessed online at [11].
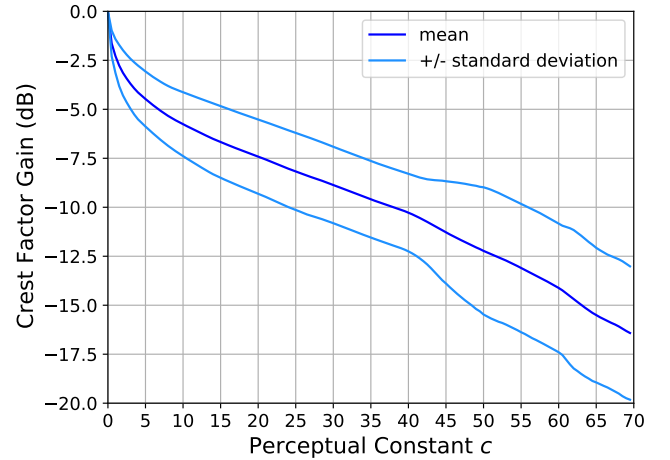
4.1. Crest Factor Gain

From a physical system design perspective, crest factor figures are commonly used to estimate loudspeakers power handling since they provide a measure of absolute peak values relative to RMS [12]. We introduced the crest factor as a suboptimal metric for increasing loudness. Fig. 1 shows the average decrease in crest factor, as computed by (2), for 16 kick drum clips taken from the dataset. It can be observed that even for small values of detectability, i.e., less than 10, the reduction can lead to a significant increase in perceived volume.

It is important to emphasize that we introduced a slight change in the matrix $\mathbf{P}_{\mathbf{x}_0}$ in all our simulations. The model from [8] assigns weights close to zero to frequencies below 30 Hz. This can cause the optimization problems in (2) and (3) to encourage solutions with energy in an inaudible fre-

quency range. Thus, we assign these weights large values so that they are effectively ignored.

Fig. 2 shows the magnitude of the spectrum of a single modified kick drum clip from the dataset for different values of the detectability parameter. For low values of this parameter, the constraint on the perceptual difference has more influence. Thus, frequencies between 100 and 200 Hz, which are less constrained by the psychoacoustic model [8], are modified the most. For larger values of detectability, reducing the absolute peak value dominates and tends to produce a line spectrum. Additionally, the solution tends to present a reduction in energy that can be easily solved by linear amplification.

Fig. 1. Crest factor reduction for the method proposed in (2) as a function of the detectability parameter c. Results are generated using the 16 clips taken from [10].
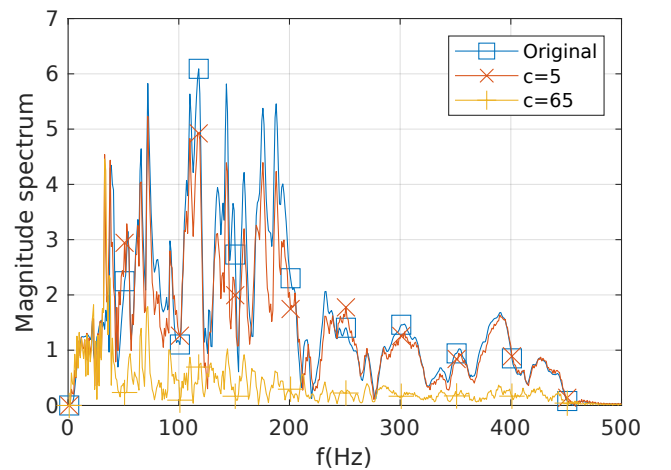
Fig. 2. Magnitude spectrum (linear) for three different values of the detectability parameter c in (2). Results generated for the clip n°1 of [11]

4.2. Loudness and Audio Quality

We compare both formulations of the method proposed, i.e., (2) and (3), against a dynamic range compressor and clipping techniques. The dynamic range compressor is implemented with an attack of 0 s, a release of 0.5 s, a ratio of 8, and a knee of 0 dB [3]. The hard clipper is easily implemented by clamping to a constant value samples above a threshold. The soft clipper is implemented by setting the appropriate parameters of a dynamic range compressor, i.e., attack and release times of 0 s, a ratio of 50, and a knee of 20dB.

In order to evaluate the performance of these methods, we use loudness and audio quality metrics. We use the Loudness Units relative to Full Scale (LUFS) [13], also known as Loudness, K-weighted, relative to full scale (LKFS)[14]. This is commonly used, for example, to comply with the audio normalization requirements of many broadcast standards. In terms of audio quality, we use, as an objective measure, the Perceptual Evaluation of Audio Quality (PEAQ) standard [15], [16]. We use 16 kick drum clips from the dataset in [10]. Note that loudness measurements specially penalize low-frequency content, thus using kick drum clips represents a challenging scenario for all these methods.

We report both loudness and objective audio quality scores as a function of peak reduction. In particular, for each of the methods evaluated, the x-coordinate values are the average, across kick drum clips, of the absolute peak value reduction in percent. Similarly, the corresponding y-coordinate values are the average, across kick drum clips, of the performance metrics, i.e., loudness or quality scores estimates. As a preprocessing stage for the loudness estimates, the samples were periodically replicated in time and truncated to 1 s. Regarding the PEAQ scores, the reference signal was processed in the same manner as for the loudness estimates. Additionally, the modified audio clips were then appropriate scaled so that they have the same loudness as the reference.

Fig. 3 shows the loudness values as a function of average peak decrease, i.e., $(||\mathbf{x}_0||_\infty - ||\mathbf{x}||_\infty)/||\mathbf{x}_0||_\infty$ in percentage values. We can observe that the hard and soft clippers outperform the rest of the methods in terms of loudness. However, as can be seen in Fig. 4, they achieve this at the expense of audio quality. Both formulations of the method proposed in this paper have similar performance, i.e., they present the highest objective score in terms of audio quality and, at the same time, they present a significant increase in loudness. These results also suggest that the dynamic range compressor presents a better tradeoff than clipping techniques in terms of loudness and audio quality.

The main difference between both formulations of the method proposed in this paper is that it is possible to either control the resulting audio quality, as in (2), or the maximum peak value of the resulting signal, as in (3). The former has the advantage of being able to provide guarantees in terms of resulting audio quality in a system design process.
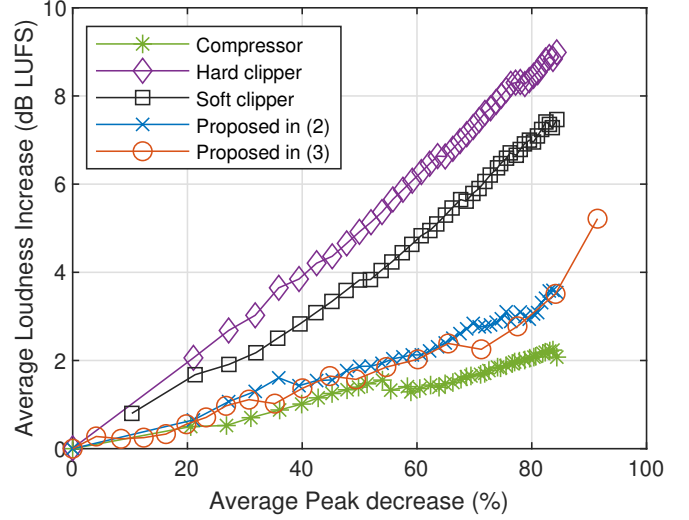
Fig. 3. Average loudness increase as a function of average decrease of the absolute peak value.
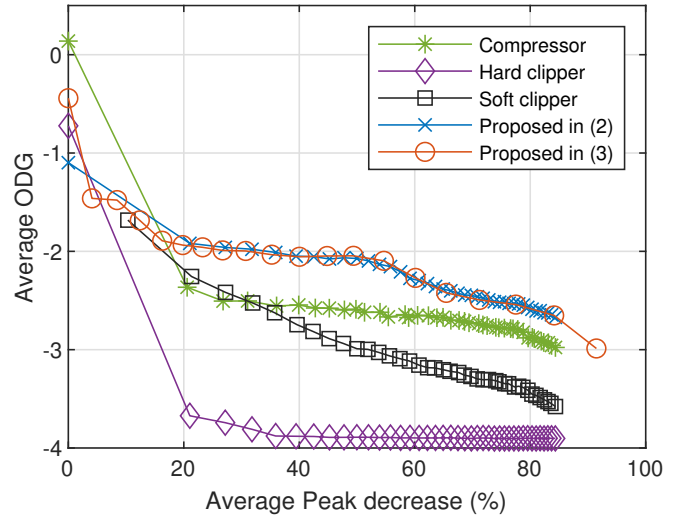
Fig. 4. Average Objective Difference Grade (ODG)—i.e., objective sound quality score estimated by PEAQ—as a function of average decrease of the absolute peak value.

5. CONCLUSION

We presented a method designed to increase loudness while reducing the absolute peak value of audio signals and, at the same time, preserving audio quality. We achieved this by exploiting a psychoacoustic model, based on multitonal masking, in order to explicitly control either the resulting audio quality or the resulting peak value depending on the formulation of the method. Our results suggest that the methods proposed present a better tradeoff between loudness and audio quality than dynamic range compressors and clipping techniques and should be validated by subjective listening test.

References

[1] R. H. Small, "Closed-box loudspeaker systems-part 1: Analysis," *Journal of the Audio Engineering Society*, vol. 20, no. 10, pp. 798–808, 1972.

[2] A. Theile, "Loudspeakers in vented boxes: Part i & ii," *J. Audio Eng. Soc*, vol. 19, no. 5,

[3] D. Giannoulis, M. Massberg, and J. D. Reiss, "Digital dynamic range compressor design–a tutorial and analysis," *Journal of the Audio Engineering Society*, vol. 60, no. 6, pp. 399–408, 2012.

[4] W. Wathen-Dunn and D. W. Lipke, "On the power gained by clipping speech in the audio band," *The Journal of the Acoustical Society of America*, vol. 30, no. 1, pp. 36–40, 1958.

[5] P. Kendrick, I. R. Jackson, F. F. Li, B. Fazenda, T. Cox, *et al.*, "Perceived audio quality of sounds degraded by non-linear distortions and single-ended assessment using hasqi," *Journal of the Audio Engineering Society*, vol. 63, no. 9, pp. 698–712, 2015.

[6] B. Defraene, T. van Waterschoot, H. J. Ferreau, M. Diehl, and M. Moonen, "Real-time perception-based clipping of audio signals using convex optimization," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 10, pp. 2657–2671, 2012.

[7] A. Beck, *First-order methods in optimization*. SIAM, 2017.

[8] S. van de Par, A. Kohlrausch, R. Heusdens, J. Jensen, and S. H. Jensen, "A perceptual model for sinusoidal audio coding based on spectral integration," *EURASIP Journal on Advances in Signal Processing*, vol. 2005, no. 9, pp. 1–13, 2005.

[9] C. H. Taal, R. C. Hendriks, and R. Heusdens, "A low-complexity spectro-temporal distortion measure for audio processing applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 5, pp. 1553–1564, 2012.

[10] T. M. team, *Sampleradar: 1,000 free drum samples*, Aug. 2020. [Online]. Available: https://www.musicradar.com/news/drums/1000-free-drum-samples.

[11] A. Jeannerot, *Dataset: "Increasing loudness in audio signals: a perceptually motivated approach to preserve audio quality"*, Jan. 2022. DOI: 10.5281/zenodo.5828375. [Online]. Available: https://doi.org/10.5281/zenodo.5828375.

[12] *Aes2-2012 aes standard for acoustics*, 2012.

[13] E. B. Union, *Recommendation r128: Loudness normalisation and permitted maximum level of audio signals*, 2010.

[14] I. T. Union, *Recommendation itu-r bs.1771-1: Requirements for loudness and true-peak indicating meters*, 2012.

[15] I. Recommendation, "1387: Method for objective measurements of perceived audio quality," *International Telecommunication Union, Geneva, Switzerland*, 2001.

[16] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, and C. Colomes, "Peaq-the itu standard for objective measurement of perceived audio quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1/2, pp. 3–29, 2000.