Perceptually Controlled Reshaping of Sound Histograms

Gaël Mahé, Mériem Jaidane

Perceptually Controlled Reshaping of Sound Histograms

Gaël Mahé* and Mériem Jaïdane

***Abstract*—Many audio processing algorithms have optimal performance for specific signal statistical distributions that may not be fulfilled for all signals. When the original signal is available, we propose to add an inaudible noise so that the distribution of the signal-plus-noise mixture is as close as possible to a given target distribution. The proposed generic algorithm (independent from the application) adds iteratively a low-power white noise to a flat-spectrum version of the signal, until the target distribution or the noise audibility is reached. The latter is assessed through a frequency masking model. Two implementations of this sound reshaping are described, according to the level of the targeted transformation and to the foreseen application: *Histogram Global Reshaping (HGR)* to change the global shape of the histogram and *Histogram Local Reshaping (HLR)* to locally "chisel" the histogram, but keeping the global shape unchanged. These two variants are illustrated by two applications where the inaudibility of the noise generated by the algorithm is required: "sparsification" for source separation, and low-pass filtering of the histogram for application of the quantization theorem, respectively. In both cases, the target histogram is reached or almost reached and the transformation is inaudible. The experiments show that the source separation performs better with HGR and that the HLR allows a better application of the quantization theorem.**

***Index Terms*—sound histogram equalization, histogram global reshaping, histogram local reshaping, noise audibility control, sparsification, histogram low-pass filtering.**

EDICS Category: AUD-SEN, AUD-AMCT, AUD-MSP

I. Introduction

HISTOGRAM equalization (HE) is well-known in image processing [1], and is mainly used to enhance the contrast of an image. This kind of processing does not belong to the traditional toolbox of sound processing, because the "contrast" of a sound is not clearly defined and may not be enhanced through simple HE.

However, some particular audio applications are improved with a histogram equalization. A correction of non-linear distortions based on HE was proposed in [2]. HE was also used as a pre-processing step in automatic speech recognition, in order to increase the robustness of the latter against noise (see for example [3], [4]). In all these works, HE was conceived as an enhancement processing, where the processed sound is aimed at being perceptually different from the original one.

Gaël Mahé is with the Laboratory of Informatics Paris Descartes (LIPADE) in Paris Descartes University, 45 rue des Saints Pères 75270 Paris cedex 06, France. e-mail: gael.mahe@mi.parisdescartes.fr - tel: +33 1 76 53 02 82

Mériem Jaïdane is with the Signals and Systems Lab (U2S) in the National School of Engineering of Tunis (ENIT), Université de Tunis El Manar, BP 37, 1002 Tunis-Belvédère, Tunisia. e-mail: meriem.jaidane@enit.utm.tn - tel: +216 71 874 700

In a different approach, the goal of HE is to enhance a specific later processing step that needs particular statistical properties, while it is essential to keep the perceptual properties of the signal unchanged. In [5], the audio signal is "Gaussianized" in order to better identify the non-linear system that will convey it. In [6]–[9], the time-frequency coefficients of audio signals are "sparsified" - *i.e.* the amount of zeros is increased [6], [7], [9] or the shape parameters of their Generalized Gaussian distributions are reduced [8] - in order to enhance source separation [7], [8] or audio-coding [6], [9]. In [10], the histogram is low-pass filtered in order to fulfill the conditions of the quantization theorem [11] and restore the histogram of the signal from that of the sub-quantized signal. In all these applications, the transformation must be perceptually transparent.

While the algorithms operating on time-frequency distributions [6]–[9] control the audibility of the transformation through perceptual models, none of the algorithms operating on time-domain samples [5], [10] provides a satisfactory perceptual control of the transformation.

Hence, our goal is to propose a generic algorithm that reshapes the histogram of the temporal samples of an audio sequence into a target histogram, under an explicit constraint of inaudibility of the additive noise generated by the transformation. The proposed algorithm is intended for any application where the original signals to be processed are available and should have a specific distribution for further optimal processing. This distribution is for example Gaussian for non-linear system identification [5], sparse for source separation [12], band-limited for application of the quantization theorem [10], adapted to the error minimization in the optimal quantization [13].

Two levels of reshaping are targeted: changing the global shape of the distribution, as in [5]–[9], or reshaping locally the histogram, as in [10], assuming that the global shape is satisfactory.

In Section II, we will present our algorithm and discuss its features. Two versions of this algorithm will be described in Sections III-A and III-B, according to the goal of the reshaping: *Histogram Global Reshaping* (HGR) and *Histogram Local Reshaping* (HLR), respectively. Finally, some experimental results with speech and music will be presented in Section IV, related to "sparsification" for a generic algorithm of blind source separation in the time-domain, and low-pass filtering of the histogram for application of the quantization theorem.

II. Principles of the algorithm

A. Framework and goals

Let x be a discrete signal of finite length N, with integer values $x(0) \dots x(N-1)$. Considering histograms whose classes are integer numbers, the histogram and the cumulative histogram of x are defined by:

$$\forall\, k \in \mathbb{Z},\ \begin{cases} f_x(k) = |\{x(n)\,|\,x(n)=k\}| \\ F_x(k) = \sum_{i=-\infty}^{k} f_x(i) = |\{x(n)\,|\,x(n)\leq k\}|, \end{cases} \tag{1}$$

respectively, where $|S|$ denotes the cardinality of a set S.

Let f_{target} be the target histogram. Our goal is to find a transformation of x into z (with integer values, too) so that:

$$\begin{cases} f_z \simeq f_{target} & (2) \\ w = z - x \text{ is inaudible with reference to } x & (3) \end{cases}$$

where w denotes the equivalent additive transformation noise. We assess its inaudibility through a frequency masking constraint based on an frequency analysis on successive frames of duration around 20 ms. Hence, reaching (2) and (3) corresponds to solving:

$$\begin{cases} \min \mathrm{d}(f_z, f_{target}) \quad \text{such that :} & (4) \\ \gamma_w(m,\nu) < \gamma_{mask}(m,\nu) \quad \forall\, \text{frame}\, m,\ \text{frequency}\, \nu & (5) \end{cases}$$

where d denotes a given distance ($\mathrm{d_{TV}}$ for total variation, $\mathrm{d_{KS}}$ for Kolmogorov-Smirnov...), $\gamma_w(m,\nu)$ denotes the power spectral density of the transformation noise w in the m^{th} frame, and $\gamma_{mask}(m,\nu)$ the masking threshold of x in the m^{th} frame, in the frequency domain. In this paper, we will consider a classical frequency-masking model [14], [15].

The proposed goal expressed by Eq. (4) and (5) exhibits two challenging constraints:

- the histogram optimization has to be performed on the whole signal, while the constraint is local and different for each frame;
- the histogram is the histogram of the time samples, while the constraints are expressed in the frequency domain.

B. State of the art

The principle of HE in [10] is to move the samples between neighboring classes of the histogram, from classes in excess to deficient classes as follows. Initially, $z = x$. For j varying from the minimum to the maximum values of x,

- If $f_z(j) - f_{target}(j) = M > 0$, M samples of z of value j are randomly selected. Each of those samples gets the value $j+1$, so that $f_z(j) = f_{target}(j)$ and $f_z(j+1) = f_z(j+1) + M$.
- If $f_z(j) - f_{target}(j) = -M < 0$, M samples of z of value $j+1$ are randomly selected. Each of those samples gets the value j. If $f_z(j+1) < M$, the missing samples are randomly selected among those of value $j+2$, and so on, until $f_z(j) = f_{target}(j)$.

At the end of this algorithm, the target histogram is exactly reached, but the algorithm does not provide any control on the audibility of the transformation. In the application that [10] dealt with, *i.e.* low-pass filtering histograms, the lower the cut-off frequency was, the higher the transformation noise w was. It was only shown on some experimental examples that w was inaudible for cut-off frequencies above a threshold depending on the signal.

The first version of HE in [5], for "Gaussianization", is based on the same algorithm as in image processing. In the time-domain the principle is to find for each sample $x(n)$ the sample $z(n)$ so that the target cumulative distribution function F_{target} in $z(n)$ matches the empirical cumulative distribution function in $x(n)$, *i.e.*:

$$F_{target}\big(z(n)\big) = F_x\big(x(n)\big) \tag{6}$$

Again, this basic algorithm makes the histogram f_z match exactly the target histogram f_{target}, but does not control the audibility of the transformation of x into z.

In order to make the transformation noise w inaudible, [5] added a constraint to (6), leading to the following optimization problem:

$$\forall\, n\,,\ \begin{cases} \min \left|F_{target}\big(z(n)\big) - F_x\big(x(n)\big)\right| & (7) \\ |w(n)| < w_{max} & (8) \end{cases}$$

where w_{max} was set so that the variance of w estimated on the whole sequence was equal to a target variance empirically chosen to ensure inaudibility. The drawback of this method is that it does not explicitly shape w according to a perceptual model, as expressed by Eq. (5) for example, which may lead to a sub-optimal trade-off between reaching the target histogram and ensuring the inaudibility of the transformation.

A perceptual control of these algorithms would require to fulfill the condition (5) on $\gamma_w(m,\nu)$. In the time domain, this spectral shaping implies introducing a particular correlation between successive values $w(n)$. This is not possible in Algorithm [10] (described at the beginning of this subsection) where the samples are not processed sequentially. In Algorithm [5], described above, Equations (7) and (8) constitute a local optimization problem, sample by sample. Replacing (8) by (5) would involve neighboring samples, which makes it difficult to keep (7) local. The proposition that follows can be seen as a generalization of [5] that takes into account the interdependence between neighboring samples induced by the frequency-domain constraint (5).

C. Proposed perceptually controlled histogram transformation

To circumvent the difficulty of taking into account the dependencies between each sample $w(n)$ and its neighbors, we propose to formulate the problem in a domain of representation where the transformation noise is white.

For this purpose, we propose a process based on the flattening/recoloring scheme [1] of Fig. 1, where T is a transformation (possibly non-linear) controlled by the distance between f_z and f_{target}, and assumed to be equivalent to the addition of a white noise. Hence, the choice of H^{-1} allows to shape in the frequency domain the transformation noise $w = z - x$ according to (5), while the specification (4) is obtained through the control of T by $\mathrm{d}(f_z, f_{target})$.

[1] Note that the flattening filter H actually flattens the masking threshold of the signal, not the signal spectrum itself, as will be seen later (see Eq. (17))

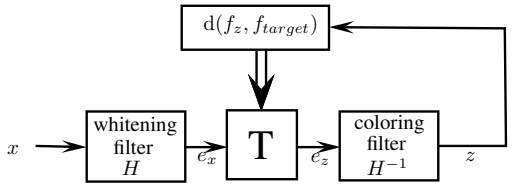

Fig. 1. Basic principle of the perceptually controlled sound histogram transformation.

We propose to implement this general scheme through the iterative algorithm illustrated by Fig. 2.

Initially,

- the original signal x is filtered by a frame-varying flattening filter H_m controlled by a psychoacoustic model of x:

$$e_x = h_m * x, \tag{9}$$

  for each m^{th} frame, where h_m is the impulse response of H_m and $*$ means discrete convolution. While the psychoacoustic model is computed on 50%-overlapping frames, the filter coefficients are defined on non-overlapping frames. Hence, each processed frame m corresponds to the central half of the samples of the m^{th} analysis frame.
- $z = x$ and $e_z = e_x$.

Then, for each iteration i (one cycle) and each discrete time n, we generate a random value $\delta w_e(i,n) \sim \mathcal{N}(0, \sigma_i^2)$. Adding $\delta w_e(i,n)$ to $e_z(n)$, the n^{th} sample of the flattened version of z, can modify $z(n) \dots z(n+L)$ (due to the coloring filter H_m^{-1}), where $L+1$ is the length of the impulse response of H_m^{-1}, denoted by h_m^{-1}, possibly infinite[2]. Denoting by $e_{z'}$ and z' the modified versions of e_z and z,

$$\forall\, k \in [0, L],\ z'(n+k) = \text{Round}\Big(\sum_{j=0}^{L} h_m^{-1}(j) e_{z'}(n+k-j) \Big) \tag{10}$$

If this modification of z reduces $\text{d}(f_z, f_{target})$, then we add $\delta w_e(i,n)$ to $e_z(n)$ and we therefore modify $z(n) \dots z(n+L)$ into $z'(n) \dots z'(n+L)$. Otherwise we do not add $\delta w_e(i,n)$ to $e_z(n)$ and z remains unchanged.

Fig. 3 presents an example of H_m^{-1}, e_z and z in the frequency domain for some frame m of a piano signal, to illustrate the flattening/coloring scheme.

This procedure is repeated several times on the whole sequence. Hence, for q iterations,

$$z = x + \text{Round}(w) \tag{11}$$

with

$$w = h_m^{-1} * w_e \tag{12}$$

where

$$w_e(n) = \sum_{i=1}^{q} \delta(i,n) \delta w_e(i,n) \tag{13}$$

where $\delta(i,n) = 0$ or 1 according to the decision of adding $\delta w_e(i,n)$ or not. All $\delta w_e(i,n)$ are independent. Assuming

[2]We will consider only causal filters.

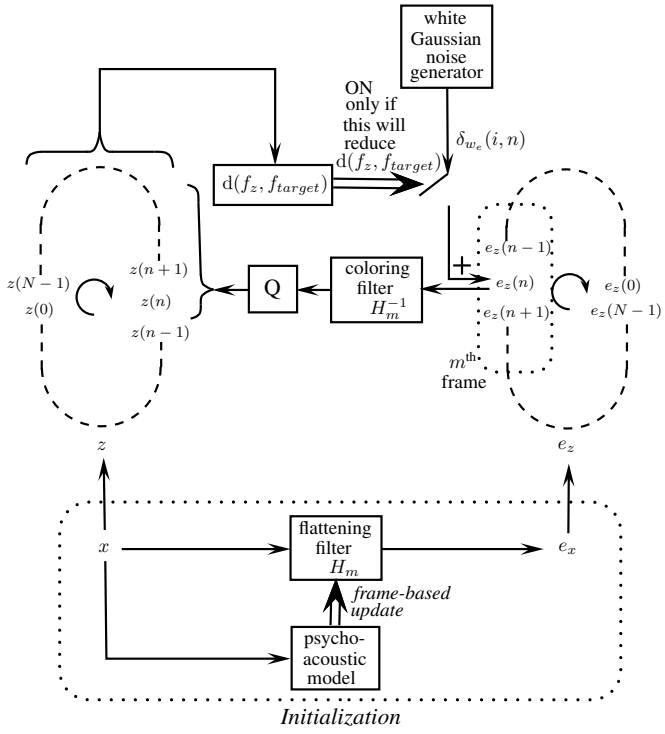

Fig. 2. Perceptually controlled sound histogram reshaping. The initialization part is performed once on the whole signal x. Then, after copying x to z and e_x to e_z, the remaining part is performed with several iterations i (↻) on the whole signals e_z and z. The "Q" box is a quantizer.

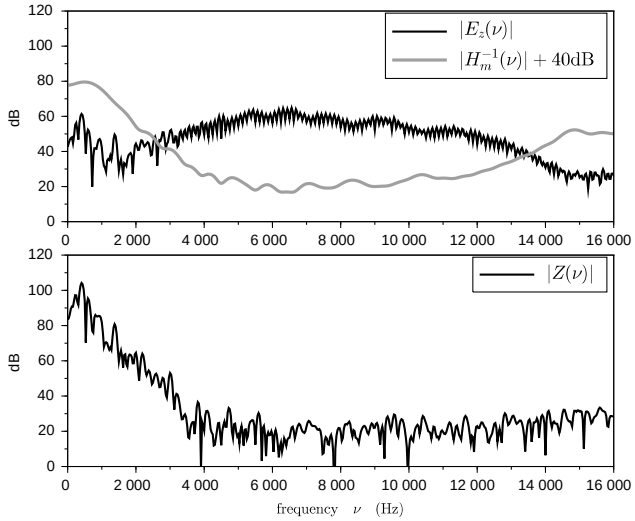

Fig. 3. For a frame m of a piano signal sampled at 32 kHz, magnitude frequency response $|H_m^{-1}(\nu)|$ of the coloring filter and magnitude spectra (in decibels) $|E_z(\nu)|$ and $|Z(\nu)|$ of signals e_z and z. The spectra are computed by Discrete Fourier Transform on 512 samples multiplied by a Hann window. The filter H_m^{-1} corresponds to the 256 central samples of this frame.

that the $\delta(i,n)$ are independent, $\delta(i,.)$ can be modeled by a Bernoulli process of parameter p_i, and then w_e is a white noise of variance:

$$\sigma_{w_e}^2 = \sum_{(i_1 \dots i_q) \in \{0;1\}^q} \Pr\big(\Lambda(n, i_1 \dots i_q)\big) \sum_{k=1}^{q} i_k \sigma_k^2 \tag{14}$$

where $\Lambda(n, i_1 \dots i_q)$ is defined in (40) (see Appendix A for

proof). In the case where $\forall\, i, \quad p_i = p$ and $\sigma_i = \sigma$,

$$\sigma^2_{w_e} = pq\sigma^2 \tag{15}$$

Hence, the constraint $\gamma_w(m,\nu) < \gamma_{mask}(m,\nu)$ (5) means:

$$|H_m^{-1}(\nu)|^2 \sigma^2_{w_e} < \gamma_{mask}(m,\nu) \tag{16}$$

which can be implemented as:

$$\begin{cases} |H_m^{-1}(\nu)|^2 = \gamma_{mask}(m,\nu) & (17) \\ \sigma_{w_e} < 1 & (18) \end{cases}$$

where the satisfaction of the constraint (18) can be controlled through the choice of q and $(\sigma_i)_{1\leq i\leq q}$.

Concerning the δ independence assumption, remind that $\delta(i,n)$ depends on the effect on $\mathrm{d}(f_z, f_{target})$ of the modification of $z(n)\dots z(n+L)$ caused by $\delta w_e(i,n)$. Since the $\delta w_e(i,n)$ are independent and the dependency of successive samples of z has only an indirect effect on the dependency of $\mathrm{d}(f_z, f_{target})$ modifications - particularly if L is high - it seems reasonable to assume the Bernoullicity of $\delta(i,.)$. The latter will be checked in the experimental part (IV) and in Appendix C. This is however not a necessary condition to get w_e stationary and white. Referring to Eq. (13), since each $\delta w_e(i,.)$ is stationary, w_e is stationary if each $\delta(i,.)$ is stationary. Concerning the whiteness, let:

$$w_e^{(i)}(n) = \delta(i,n)\delta w_e(i,n) \tag{19}$$

The auto-correlation of $w_e^{(i)}$ is defined by:

$$\begin{aligned} &\Gamma_{w_e^{(i)}}(k) \qquad (20)\\ &= \mathrm{E}[\delta(i,n)\delta w_e(i,n)\delta(i,n+k)\delta w_e(i,n+k)] \\ &= \Pr\big(\delta(i,n)\delta(i,n+k)=1\big) \\ &\quad \times \mathrm{E}[\delta w_e(i,n)\delta w_e(i,n+k) \,|\, \delta(i,n)\delta(i,n+k)=1] \end{aligned}$$

Hence, w_e is white if for each i and for $k \neq 0$:

$$\mathrm{E}[\delta w_e(i,n) \,|\, \delta(i,n)=1] = 0 \tag{21}$$

$$\mathrm{E}[\delta w_e(i,n)\delta w_e(i,n+k) \,|\, \delta(i,n)\delta(i,n+k)=1] = 0 \tag{22}$$

The latter properties (21,22) will be checked in the experimental part (IV) and in Appendix C.

The algorithm is repeated until f_z is close enough to f_{target} or the estimated σ_{w_e} ($\hat{\sigma}_{w_e} = \frac{1}{N}\sum_{n=0}^{N-1} w_e(n)^2$) reaches 1. The detailed algorithm is presented in Fig. 4. Running several iterations with small values of σ_i instead of few iterations with $\sigma_i \simeq 1$ allows to control a slow increase of σ_{w_e} until 1, avoiding reaching 1 if not necessary. Moreover, a modification of a sample $e_z(n)$ that was not possible at an iteration may become possible at the next iteration, since the modifications of the other samples that occurred meanwhile changed the histogram. Hence, small variations distributed among all the samples can achieve the same histogram reshaping as larger variations concentrated on fewer samples, with a lower variance of the transformation noise w_e.

To decide if a sample modification will make the histogram f_z closer to the target histogram f_{target}, we use a distance D (see Fig. 4) that can be different from the distance d. In particular, a distance d based on the L_∞-norm is unlikely modified by a local modification of f_z, although

$z \leftarrow x$, $f_z \leftarrow f_x$, $F_z \leftarrow F_x$ and $e_z \leftarrow e_x$
Fix ε, ε', MAX_IT, Δ_d^{min}, $i=0$, $(\sigma_\ell)_{1\leq\ell\leq MAX\_IT} < 1$
repeat
  $i \leftarrow i+1$
  for $n = 0 \to N-1$ **do**
    $z' \leftarrow z$, $f_{z'} \leftarrow f_z$, $F_{z'} \leftarrow F_z$ and $e_{z'} \leftarrow e_z$
    Generate $\delta w_e \sim \mathcal{N}(0, \sigma_i^2)$
    $e_{z'}(n) \leftarrow e_{z'}(n) + \delta w_e$
    $\Delta D \leftarrow 0$
    for $k = 0 \to L$ **do**
      $z'(n+k) \leftarrow \mathrm{Round}\big(\sum_{j=0}^{L} h_m^{-1}(j) e_{z'}(n+k-j)\big)$
      if $z'(n+k) \neq z(n+k)$ **then**
central block
        Update $f_{z'}$
        Compute $\Delta D = \mathrm{D}(f_{z'}, f_{target}) - \mathrm{D}(f_z, f_{target})$
      end if
    end for
    if $\Delta D \leq 0$ **then**
      $z \leftarrow z'$, $f_z \leftarrow f_{z'}$, $F_z \leftarrow F_{z'}$ and $e_z \leftarrow e_{z'}$
      $w_e(n) \leftarrow e_z(n) - e_x(n)$
    end if
  end for
until $\mathrm{d}(f_z, f_{target}) < \varepsilon \vee \frac{1}{N}\sum_{n=0}^{N-1} w_e(n)^2 > 1-\varepsilon' \vee i > MAX\_IT \vee \Delta D < \Delta_d^{min}$

Fig. 4. Perceptually controlled histogram transformation. D is a distance that can be different from d. The central block depends on the application case: *HGR* (see Fig. 6) or *HLR* (see Fig 7).

the latter can contribute to make f_z closer to f_{target} and thus reduce $\mathrm{d}(f_z, f_{target})$ in the long term. The distance D will be specified in Section III.

Each step (i,n) of the algorithm causes either nothing, or simultaneously a decrease of $\mathrm{D}(f_z, f_{target})$ and an increase of $\hat{\sigma}_{w_e}$. If the convergence of D is equivalent to that of d, this makes the algorithm converge as long as it is possible to modify one sample $e_z(n)$ so that $\mathrm{D}(f_z, f_{target})$ decreases. The stop conditions of the loop however make the algorithm stop if the decrease of d during one iteration on the whole signal is too small.

The coefficients of the filter H_m are updated at the same rate as the masking threshold of the signal x. Since the amount of updates of $z(n)\dots z(n+L)$ and f_z for each n depends on L, H_m^{-1} should preferably be an FIR filter. The shorter its impulse response, the rougher the approximation of the masking threshold by $|H_m^{-1}(\nu)|^2$ (Eq. (17)). Consequently, a trade-off must be done between complexity and perceptual accuracy. The global complexity of the algorithm will we calculated in Section III.

Within this general frame, one must specify:

- the psychoacoustic model and its approximation through the frequency response $|H_m^{-1}(\nu)|$, which condition the perceptual control and the complexity of the algorithm;
- the number of iterations q and the standard deviations $(\sigma_i)_{1\leq i\leq q}$, which control the convergence;
- the distances d and D, which define how the similarity between histograms is measured.

The latter will be specified in Section III, while the former two will be set in the experimental section (IV).

D. Perceptual issues

The coefficients of H_m^{-1} are updated every frame, in order to match the psychoacoustic model. The phase of this filter may change, which may cause discontinuities in the output signal at some of the frame transitions. Hence, the transformation noise w may occasionally have audible discontinuities (clicks) that are not taken into account by the psychoacoustic model, since they originate from the model update itself.

To avoid this, at the end of each iteration of the algorithm, we (i) extract the transformation noise $w = z - x$; (ii) low-pass-filter its discontinuities that occur at frame transitions and that are greater than a locally adapted threshold; (iii) add this smoothed version of w to the original signal x, thus yielding a new z. The length of the impulse response of the low-pass filter must be short enough to ensure a limited impact on P_z. The details of the smoothing process are specified in Appendix B.

Additionally, as shown in [16], fulfilling the inaudibility constraint (5), where γ_{mask} is provided by a frequency-masking model, ensures the inaudibility of the transformation noise, but only in average for a given set of audio signals. Actually, the performance of the model depend on the signal, so that γ_{mask} should be multiplied by an attenuation factor depending on the audio content.

Consequently, for each specific audio signal, all filters H_m are multiplied by the same attenuation factor λ chosen in such a way that a white Gaussian noise of variance 1 filtered by the successive λH_m^{-1} and added to the audio signal is inaudible.

III. Implementations of the Histogram Reshaping Algorithm

Two different goals may be assigned to this general algorithm: changing the global shape of the histogram (as in [5] and [8]) or, assuming the global shape is satisfying, matching accurately a target histogram (as in [10]). Fig. 5 illustrates these two transformations: in the first case (Fig. 5a), we focus on the global shape of the histogram, while in the second case (Fig. 5b), the histogram has to be locally "chiseled", keeping the global shape unchanged. We call these two tasks *Histogram Global Reshaping (HGR)* and *Histogram Local Reshaping (HLR)*, respectively. The behavior of the previously described algorithm will be oriented through the choice of the distance used to measure the dissimilarity between two histograms (respectively the Kolmogorov-Smirnov distance and the total variation distance, as will be detailed later), in two specific implementations.

A. Histogram Global Reshaping (HGR)

To change the global shape of the histogram, each step of the algorithm must reduce the difference between F_z and F_{target}, respectively the cumulative histogram of z and the target cumulative histogram. The interest of using the cumulative histogram is the underlying integration of the histogram, which smooths the local differences between f_z and f_{target}. Hence, sample modifications directed by $F_z - F_{target}$ help focusing on the global shape of the histogram.

For this purpose, the convergence will be assessed through

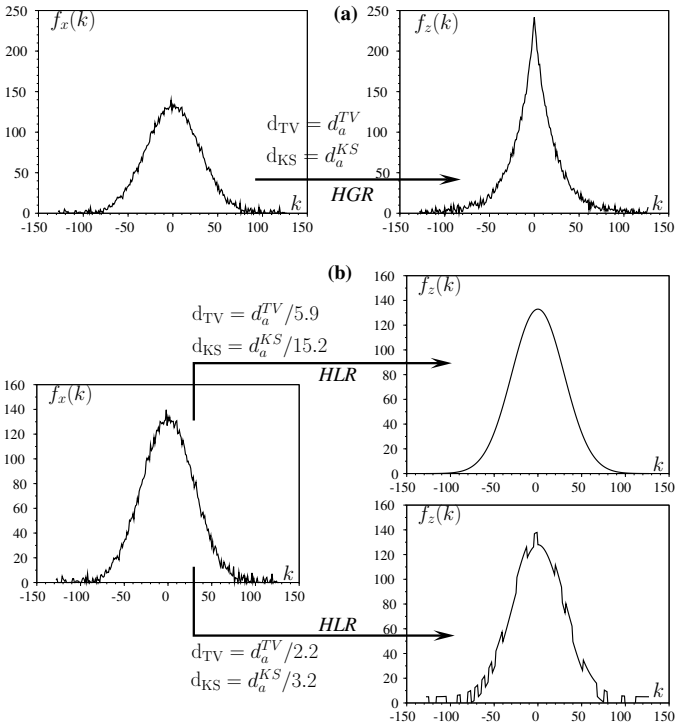

Fig. 5. Examples of Histogram Global Reshaping (HGR, a) and Histogram Local Reshaping (HLR, b). The total variation distance ($\mathrm{d_{TV}}$) is more sensitive to local differences than the Kolmogorov-Smirnov distance ($\mathrm{d_{KS}}$).

the Kolmogorov-Smirnov distance[3] and we define the distance D by:

$$\mathrm{D}(f_z, f_{target}) = \sum_{\mathbb{Z}} \big(F_z - F_{target}\big)^2. \tag{23}$$

Since D is sensitive to any sample modification, it is used to decide to add (or not) $\delta_{w_e}(i,n)$ to $e_z(n)$, while the Kolmogorov-Smirnov distance, sensitive to global modifications (L_∞ norm), is used only at the end of each cycle of the algorithm.

For each iteration i and each discrete time n, adding $\delta w_e(i,n) \sim \mathcal{N}(0,\sigma^2)$ to $e_z(n)$ modifies $z(n)\dots z(n+L)$ into $z'(n)\dots z'(n+L)$. Referring to (10), each $z(n+k)$ with $0 \le k \le L$ is modified by approximately $h^{-1}(k)\delta w_e(i,n)$. If

$$\begin{array}{rcl} F_z(z(n+k)-1) & < & F_{target}(z(n+k)-1) \\ \big(\text{resp.}\quad F_z(z(n+k)-1) & > & F_{target}(z(n+k)-1)\big) \end{array}$$

a negative (resp. positive) value $h^{-1}(k)\delta w_e(i,n)$ tends to make F_z closer to F_{target}.

Let us define the integer interval I_k:

$$I_k = \begin{cases} [z(n+k), z'(n+k)[& \text{if } z(n+k) \le z'(n+k) \\ [z'(n+k), z(n+k)[& \text{if } z'(n+k) < z(n+k) \end{cases} \tag{24}$$

For each k from 0 to L, $\mathrm{D}(f_z, f_{target})$ varies by:

$$\Delta_k D = \sum_{I_k} \big(F_{z'}^{(k)} - F_{target}\big)^2 - \big(F_{z'}^{(k-1)} - F_{target}\big)^2, \tag{25}$$

[3]Let f and g be two histograms and F and G the respective corresponding cumulative histograms, as defined in Section II. The Kolmogorov-Smirnov distance is defined by: $\mathrm{d_{KS}}(f,g) = \frac{1}{N}\sup_{k\in\mathbb{Z}}|F(k) - G(k)|$.

$f_{z'}(z'(n+k)) \leftarrow f_{z'}(z'(n+k)) + 1$
$f_{z'}(z(n+k)) \leftarrow f_{z'}(z(n+k)) - 1$
if $z'(n+k) < z(n+k)$ **then**
  $I_k = [z'(n+k), z(n+k)[$
  $\Delta_k D^- = \sum_{I_k} |F_{z'} - F_{target}|^2$
  In I_k, $F_{z'} \leftarrow F_{z'} + 1$
else
  $I_k = [z(n+k), z'(n+k)[$
  $\Delta_k D^- = \sum_{I_k} |F_{z'} - F_{target}|^2$
  In I_k, $F_{z'} \leftarrow F_{z'} - 1$
end if
$\Delta D \leftarrow \Delta D + \sum_{I_k} |F_{z'} - F_{target}|^2 - \Delta_k D^-$

Fig. 6. Central block of perceptually controlled histogram transformation (see Fig 4) for *Histogram Global Reshaping*.

where $F_{z'}^{(k)}$ is the new cumulative histogram after modifying $z(n+k)$ (with $F_{z'}^{(-1)} = F_z$).

If the total variation of $\mathrm{D}(f_z, f_{target})$ resulting from the $L+1$ modifications of z is negative, *i.e.*

$$\Delta D = \sum_{k=0}^{L} \Delta_k D \leq 0, \tag{26}$$

we carry out these modifications of $e_z(n)$ and $z(n) \dots z(n+k)$, otherwise we cancel them (see Fig. 4 with $\mathrm{d} = \mathrm{d_{KS}}$ and central block detailed in Fig. 6).

Note that a convergence according to $\mathrm{D}(f_z, f_{target})$ (L_2 norm) means a convergence according to $\mathrm{d_{KS}}(f_z, f_{target})$ (L_∞ norm)). Hence, reducing $\mathrm{D}(f_z, f_{target})$ contributes, in the long term, to a decrease of $\mathrm{d_{KS}}(f_z, f_{target})$.

B. Histogram Local Reshaping (HLR)

For a fine adjustment of a histogram having the same global shape as f_{target}, samples of z have to move locally from values in excess in the histogram to deficient values. We assess the convergence through the total variation distance[4] ($\mathrm{d_{TV}}$), which is sensitive to local differences between histograms.

For each i, n, adding $\delta w_e(i,n) \sim \mathcal{N}(0, \sigma^2)$ to $e_z(n)$ modifies $z(n) \dots z(n+L)$ into $z'(n) \dots z'(n+L)$. Each of these $L+1$ modifications causes a variation of $\mathrm{d_{TV}}(f_z, f_{target})$ in $\{0; \pm\frac{1}{N}\}$, according to the relative values of $f_{target}(z'(n+k))$ and $f_z(z'(n+k))$ on the one hand, of $f_{target}(z(n+k))$ and $f_z(z(n+k))$ on the other (see Table I). If the global variation of $\mathrm{d_{TV}}$ resulting from the $L+1$ modifications of z is negative then we carry out these modifications, otherwise we cancel them.

The detailed algorithm is presented in Fig. 4 with $\mathrm{D} = \mathrm{d} = \mathrm{d_{TV}}$ and central block detailed in Fig. 7. This algorithm ensures the decrease of $\mathrm{d_{TV}}(f_z, f_{target})$.

C. Computational complexity

From Fig. 4, the mean algorithmic complexity C after the initialization phase, measured by the number of multiplications-accumulations (MAC), is approximated as follows:

$$C \simeq N_{it} N (L+1) \Big(L + 1 + \Pr\big(z'(n) \neq z(n)\big) C' \Big), \tag{27}$$

where N_{it} and C' denote the number of iterations and the mean complexity of the central block, respectively. Referring to Fig. 6 and 7,

$$C' = 3\mathrm{E}[|z'(n) - z(n)||z'(n) \neq z(n)] + 4 \quad \text{for HGR} \tag{28}$$
$$2 \leq C' \leq 4 \quad \text{for HLR} \tag{29}$$

We empirically found the expectation in Eq. 28 equal to 1.5 to 3, and $\Pr\big(z'(n) \neq z(n)\big) \simeq 0.5$, so that Eq. 27 becomes:

$$C \simeq N_{it} N (L+1)(L+\kappa), \tag{30}$$

with $2 \leq \kappa \leq 7.5$. Although the complexity is linear in N, the multiplicative factor is quadratic in L, so that flattening filters with long impulse responses can slow down the execution.

[4]Let f and g be two histograms as defined in Section II. The total variation distance is defined by: $\mathrm{d_{TV}}(f,g) = \frac{1}{2N} \sum_{k \in \mathbb{Z}} |f(k) - g(k)|$.

TABLE I
POSSIBLE VARIATIONS OF $\mathrm{d_{TV}}(f_z, f_{target})$. $n_k = n + k$

	$f_{target}(z'(n_k)) > f_z(z'(n_k))$	$f_{target}(z'(n_k)) \leq f_z(z'(n_k))$
$f_{target}(z(n_k)) \geq f_z(z(n_k))$	0	$+1/N$
$f_{target}(z(n_k)) < f_z(z(n_k))$	$-1/N$	0

if $f_{z'}(z(n+k)) > f_{target}(z(n+k))$ **then**
  $\Delta D \leftarrow \Delta D - \frac{1}{2N}$
else
  $\Delta D \leftarrow \Delta D + \frac{1}{2N}$
end if
if $f_{z'}(z'(n+k)) < f_{target}(z'(n+k))$ **then**
  $\Delta D \leftarrow \Delta D - \frac{1}{2N}$
else
  $\Delta D \leftarrow \Delta D + \frac{1}{2N}$
end if
$f_{z'}(z'(n+k)) \leftarrow f_{z'}(z'(n+k)) + 1$
$f_{z'}(z(n+k)) \leftarrow f_{z'}(z(n+k)) - 1$

Fig. 7. Central block of perceptually controlled histogram transformation (see Fig 4) for *Histogram Local Reshaping*.

IV. EXPERIMENTAL RESULTS

The algorithm (with its two variants HGR and HLR) was tested on music (instruments and singing voice) and on speech. The signals have integer values in $[-2^{15}; 2^{15}-1]$.

In the following experiments, the filters H_m^{-1} approximating the masking threshold (17) are FIR filters of orders L. The coefficients of H_m^{-1} are those of the auto-regressive model of the inverse of the masking threshold. Indeed, considering $1/\gamma_{mask}(m,\nu)$ as a power spectral density, one can compute the corresponding auto-correlation coefficients by inverse DFT. Then we derive from these coefficients an auto-regressive model of order L corresponding to a transfer function of form $\sigma_m/(1 + \sum_{i=1}^{L} a_i^{(m)} z^{-i})$ (see [17], Chap. 11). The corresponding frequency response is a smooth version of $1/\gamma_{mask}(m,\nu)$. Hence, the frequency response of $H_m^{-1}(z) = (1 + \sum_{i=1}^{L} a_i^{(m)} z^{-i})/\sigma_m$ approximates the masking threshold.

The choice of L depends on the sampling frequency and the shape of the masking threshold. In Subsection IV-A, where music signals sampled at 32kHz are processed, we chose $L = 50$. In Subsection IV-B, where narrow-band speech is processed, the masking threshold is smoother than for music at higher sampling frequency, so that $L = 10$ is sufficient (reminding that increasing L increases the complexity).

To set the attenuation factor λ applied to all H_m^{-1}, mentioned in Subsection II-D, the masking was assessed through:

- the Objective Difference Grade (ODG) for music, predicted by PQevalAudio[5] according to PEAQ [18], which provides values between -4 and 0. A value greater than -1 indicates an inaudible distortion. We fixed λ for the ODG to be -0.5 ± 0.1.
- the Mean Opinion Score (MOS) for speech, estimated by PESQ [19], which provides scores between -0.5 and 4.5. An estimated MOS greater than 4 indicates an inaudible distortion. We fixed λ for the MOS to be 4.1 ± 0.1.

We present one application for each algorithm. Video and audio demonstrations, as well as the Scilab codes of the algorithms, are available at `http://up5.fr/SoundHistograms`

A. Histogram Global Reshaping for "Sparsification"

We ran the *HGR* algorithm on various musical tracks of the QUASI database [20], [21] sampled at 32 kHz. The masking threshold was updated every 8ms according to the MPEG-1 first model [15], computed on 50%-overlapping frames of length 512 (16ms), apodized by a Hann window. We set for σ_i a constant value of 0.25.

Here, we present the results for an 11s sequence of bass and a 10 s sequence of piano. These sequences have generalized Gaussian (GG) distributions[6] of shape parameters 1.6 and 2.1, respectively (estimation by the moment method [22]). We propose to divide their shape parameters by 2 by running the *Histogram Global Reshaping* algorithm (Fig. 6) with target histograms f_{target} matching GG distributions of shape parameters 0.8 and 1.05, respectively.

Fig. 8 presents the convergence of the algorithm in terms of Kolmogorov-Smirnov distance, in parallel with the estimation of σ_{w_e}, for both the bass sequence (8a) and the piano sequence (8b). For the bass, the algorithm reaches the bound of the inaudibility constraint ($\sigma_{w_e} < 1$) just at the convergence point, while in the case of the piano, the bound of the inaudibility constraint is reached several iterations before reaching the target distribution. In each iteration, each second of signal needs 90 to 120 s to be processed, with our Scilab implementation on a standard computer.

The assumption that each $\delta(i,.)$ can be modeled by a Bernoulli process was validated by statistical tests detailed in Appendix C.

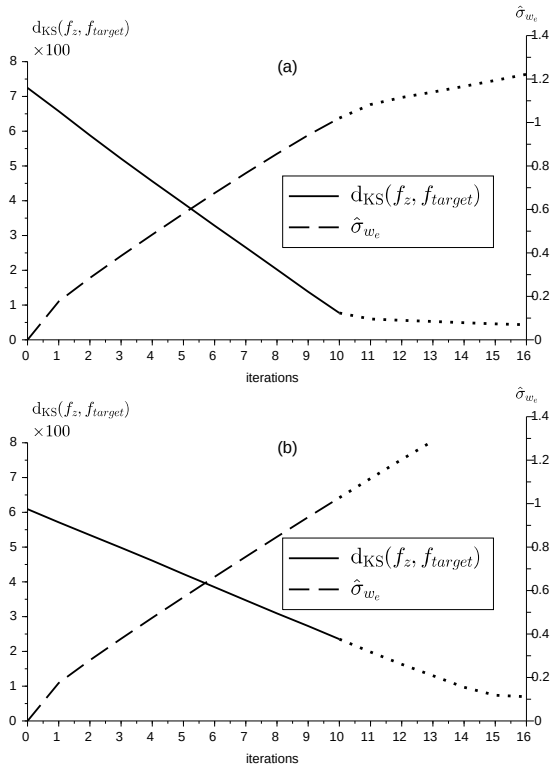

Fig. 8. *Histogram Global Reshaping* for an 11s bass sequence (a) and a 10s piano sequence (b). Initial shape parameters 1.6 and 2.1 respectively. Target shape parameters 0.8 and 1.05 respectively. Evolution of the estimated standard deviation of w_e ($\hat{\sigma}_{w_e}$) and of the Kolmogorov-Smirnov distance to the target distribution ($\mathrm{d_{KS}}(f_z, f_{target})$). The dotted lines correspond to the prolongation of the algorithm beyond its normal stop-point (when $\hat{\sigma}_{w_e}$ reaches 1).

Fig. 9 compares the cumulative histograms F_x, F_z and F_{target} at the stop-point of the algorithm (end of the 10^{th} iteration) for bass and piano (9a and 9b, respectively). For the bass (Fig. 9a), the global shape of f_z matches that of f_{target}, which leads to very close cumulative histograms. The histogram f_z has a final estimated shape parameter of 1.2[7]. As indicated by Fig. 9b, the final histogram for piano is intermediate between the original one and the target one. The transformed signal z has an estimated shape parameter of 1.7.

The inaudibility of the transformation was assessed through the Objective Difference Grade (ODG), predicted by PQevalAudio. We obtained an ODG of -0.9 for the bass, and -1.0 for the piano (inaudible distortions).

To compare the performance of this algorithm to that of [5], we plotted $\mathrm{d_{KS}}(f_z, f_{target})$ as a function of the ODG for both algorithms, for the previous bass and piano signals. As indicated by Fig 10, the HGR algorithm reaches a much better trade-off between audio quality and histogram reshaping.

A possible application of this sparsification is source separation in time domain, since blind source separation algorithms (BSS) are known to reach better performance with sparse

[5] `http://www-mmsp.ece.mcgill.ca/Documents/Software/Packages/AFsp/PQevalAudio.html`
Note that PQevalAudio is calibrated to signals sampled at 48 kHz, while our signals are sampled at 32 kHz. Consequently, we up-sampled them to 48 kHz before measurement

[6] The probability density function of a generalized Gaussian distribution is of the form: $f(x) = \beta/\big(2\alpha\Gamma(1/\beta)\big) \exp\big(-(|x-\mu|/\alpha)^{\beta}\big)$, where μ, α, and β denote the mean, scale and shape parameters, respectively. $\beta = 2$ for a Gaussian distribution. Γ denotes the gamma function, defined for each $z \in \mathbb{C} \setminus \mathbb{Z}$ by: $\Gamma(z) = \int_0^{\infty} t^{z-1}\mathrm{e}^{-t}\mathrm{d}t$

[7] Since the final Kolmogorov-Smirnov distance, 8×10^{-3}, is very weak, the shape parameter estimation method may not be accurate

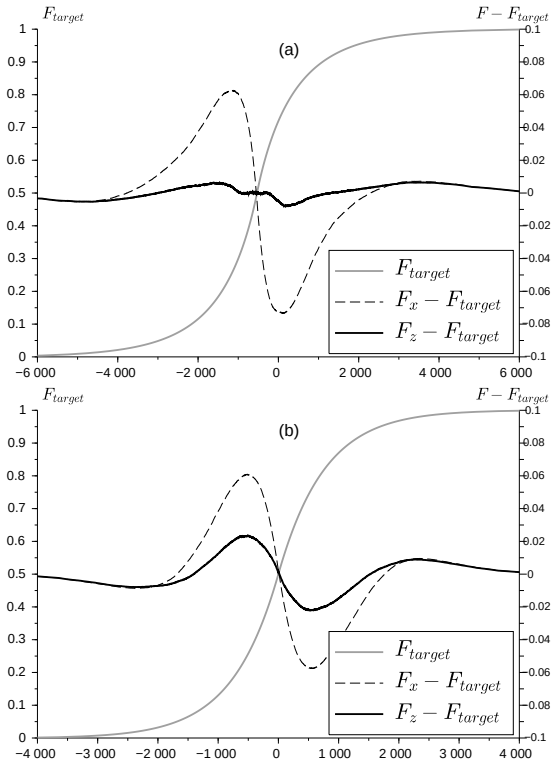

Fig. 9. Target cumulative histogram F_{target}, difference between the original cumulative histogram F_x and F_{target}, difference between the cumulative histogram F_z of the "sparsified" signal z and F_{target}, for an 11s bass sequence (a) and a 10s piano sequence (b).

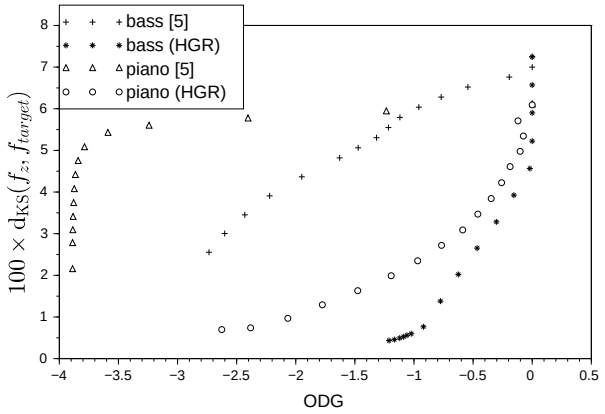

Fig. 10. Kolmogorov-Smirnov distance to the target histogram, $\mathrm{d_{KS}}(f_z, f_{target})$, as a function of the ODG, for the bass and piano signals, both processed by the algorithm [5] and by the HGR algorithm. For [5], each point corresponds to a given value of w_{max}. For HGR, each point corresponds to the result of an iteration.

signals — *i.e.* with lower shape parameters in the case of generalized Gaussian distribution — and to behave badly if the distribution is close to a Gaussian (see for example [12], [23]). Our proposal is applicable to contexts where the original sources are available before mixing, like informed source separation (ISS[8]) [24].

To illustrate this proposal, we compared the performance of a classic BSS algorithm on stereo mixtures of pairs of instruments (among singing voice, piano, acoustic guitar, electric guitar, and keyboards), sparsified according to our algorithm, to the performance obtained without sparsification. The experiment was run as illustrated by Fig. 11. We sparsified each source signal according to the *HGR* algorithm, with a target shape parameter equal to half of that of the original distribution of the signal. We chose as mixture matrix:

$$\mathbf{A} = \begin{pmatrix} 1/3 & 2/3 \\ 2/3 & 1/3 \end{pmatrix} \tag{31}$$

We used the Scilab implementation[9] of the FastICA algorithm [25] to separate the sources.

Since the Independent Component Analysis (ICA) of two mixtures of two sources is the easiest task of BSS, the sources are often well separated without any pre-processing. Consequently we focused the experiment on mixtures badly separated by FastICA. We ran this experiment on four duos from the QUASI database. For each of them, we evaluated:

- the sparsity of the original and sparsified signals, through the shape parameter β of the distribution;
- the quality of the sparsified mix, compared to the original one, through the ODG (predicted by PQevalAudio);
- the performance of the source separation, through the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Source to Artifact Ratio (SAR) as defined by [26].

The results are summarized in Table II and Fig 12[10] and can be heard on the aforementioned web page. For mixtures A, B and D, the sparsification clearly enhances the performance of source separation. Note that the SIR and SDR increase even if the shape parameter β is not much reduced (see Mixtures C and D). For mixtures that are already fairly separated without sparsification, like Mixture C, the sparsification leads to a lower enhancement of separation. In Experiment B, since the shape parameter of the guitar signal is around 3, sparsifying this signal could make its distribution close to a Gaussian one. Consequently, we let it unsparsified.

The ODG of the mixes (Table II) and the SxRs (Fig. 12) measured separately the perceptual impairments of the mixes due to sparsification and the separation enhancement due to sparsification, respectively. To measure the global gain due to sparsification, encompassing both effects, we computed the perceptual scores provided by the PEASS toolkit [27], using the original signals (x_1 and x_2 in Fig. 11) as reference signals for both couples of separated signals: $(\hat{x}_1, \hat{x}_2)$ and $(\hat{z}_1, \hat{z}_2)$. PEASS provides four perceptual scores (PS): overall (OPS),

[8]The typical application case of ISS is the following: musical sources are recorded separately in a studio on different tracks; the tracks are watermarked before mixing; then, any final user can separate the sources from the mixture by using the information conveyed by the watermark, assuming the latter has been specifically designed for this purpose.

[9]`http://research.ics.aalto.fi/ica/fastica/`

[10]In some cases, the separation performed by FastICA is variable. Hence, we ran the separation algorithm 20 times for each mixture and displayed the means and the confidence intervals (if applicable) of the SxR values.

TABLE II
RESULTS OF SPARSIFICATION OF THE MIXTURES.

Mixture	instruments	shape parameter	ODG mix
A	voice piano	$1.8 \rightarrow 1.6$ $2.1 \rightarrow 1.7$	-0.8
B	guitar keyboards	3 $1.8 \rightarrow 1.3$	-0.4
C	voice 1 voice 2	$1.1 \rightarrow 1.0$ $1.0 \rightarrow 0.9$	-0.6
D	guitar solo guitar acoustic	$1.5 \rightarrow 1.5$ $1.4 \rightarrow 1.1$	-0.8

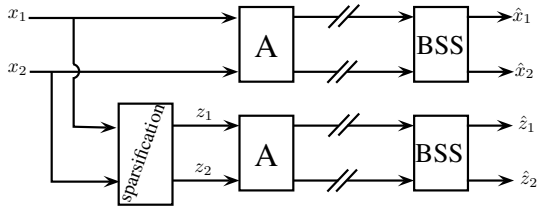

Fig. 11. Blind source separation (BSS) of a stereo mixture (mixture matrix A), without (top) and with (bottom) sparsification of the source signals.

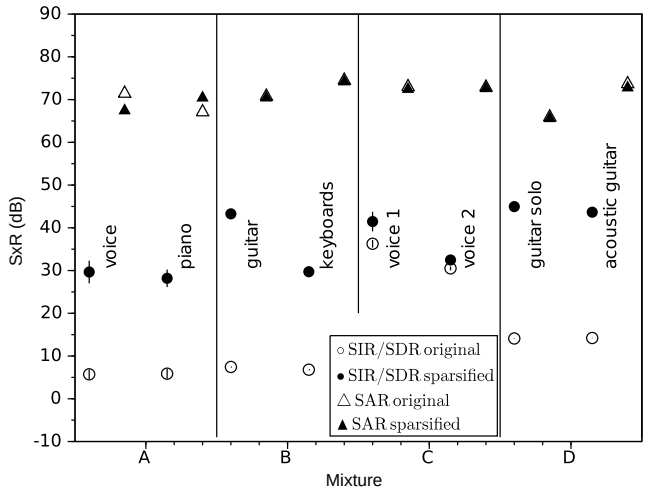

Fig. 12. Performance of source separation with and without sparsification for 4 mixtures A, B, C, and D. Vertical bars represent confidence intervals (when applicable). In Mixture B, the guitar signal was not sparsified. SIR = Source to Interference Ratio; SDR = Source to Distortion Ratio; SAR = Source to Artifact Ratio.

target-related (TPS), interference-related (IPS), and artifacts-related (APS). These scores are summarized in Fig. 13 for the four considered mixes. The same observations as for Fig. 12 can be made. Hence, when taking as reference the unsparsified signals, the sparsification-mixing-unmixing process leads to better separation quality than mixing-unmixing.

B. Low-pass filtering of the histogram

We will illustrate the *Histogram Local Reshaping* algorithm in the case of low-pass filtering of the histogram for the application of the quantization theorem.

Let x be a discrete-valued signal with (discrete) probability density function (PDF) f_x. The characteristic function of x

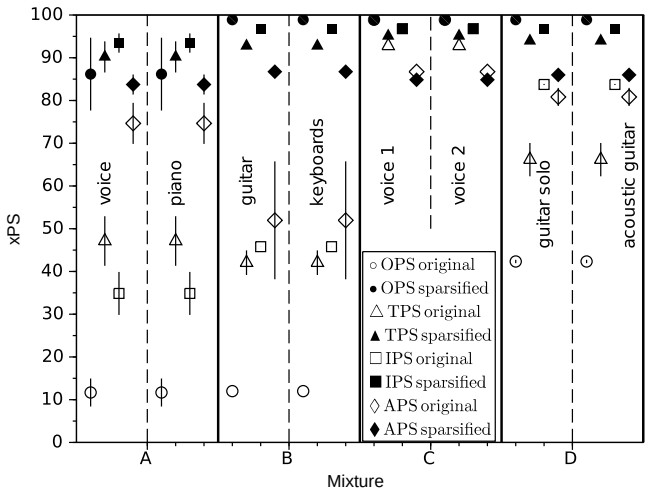

Fig. 13. Global perceptual impact of sparsification on source separation for 4 mixtures A, B, C, and D. OPS = Overall Perceptual Score, TPS = Target-related Perceptual Score, IPS = Interference-related Perceptual Score, APS = Artifacts-related Perceptual Score. Vertical bars represent confidence intervals (when applicable).

is defined by $\mathrm{DFT}^{-1}(f_x)$. According to the sub-quantization theorem [10], adapted from the quantization theorem [11] to discrete-valued signals, if the characteristic function of x is equal to zero for frequencies $|\nu| > \frac{1}{2K}$ in $[-\frac{1}{2}; \frac{1}{2}]$, then the probability density function of x can be derived from that of the signal x_Q resulting from the sub-quantization of x with a factor K.

The original PDF is recovered through filtering the PDF of x_Q by a filter of frequency response $G(\nu)$:

$$G(\nu) = \begin{cases} \frac{1}{R(\nu)} & \text{if } |\nu| < \frac{1}{2K} \\ 0 & \text{otherwise} \end{cases} \tag{32}$$

where:

$$R(\nu) = \begin{cases} K & \text{if } \nu \in \mathbb{Z} \\ \frac{sin(\pi K \nu)}{sin(\pi \nu)} \exp(j\pi\nu) & \nu \notin \mathbb{Z} \text{ and } K \text{ even} \\ \frac{sin(\pi K \nu)}{sin(\pi \nu)} & \nu \notin \mathbb{Z} \text{ and } K \text{ odd} \end{cases} \tag{33}$$

The same results hold for histogram instead of PDF.

Hence, to fulfill the condition of the sub-quantization theorem, the histogram must be low-passed, with a cut-off frequency $1/2K$. Since this may not change the global shape of the histogram, the *HLR* algorithm is appropriate for this purpose. The *HLR* algorithm was run on speech signals from the TIMIT database [28], sampled at 8kHz. The masking threshold was approximated by the power spectral density of the signal, minus a given offset [29], computed on 50%-overlapping frames of length 256 (32ms), apodized by a Hamming window. We set $K = 16$, so that the target histogram is the result of the low-pass filtering of f_x with a normalized cut-off frequency 1/32.

We present the results for a 3s sentence pronounced by a female and a male speakers. Fig. 14 illustrates the convergence of the algorithm in terms of total variation distance, in

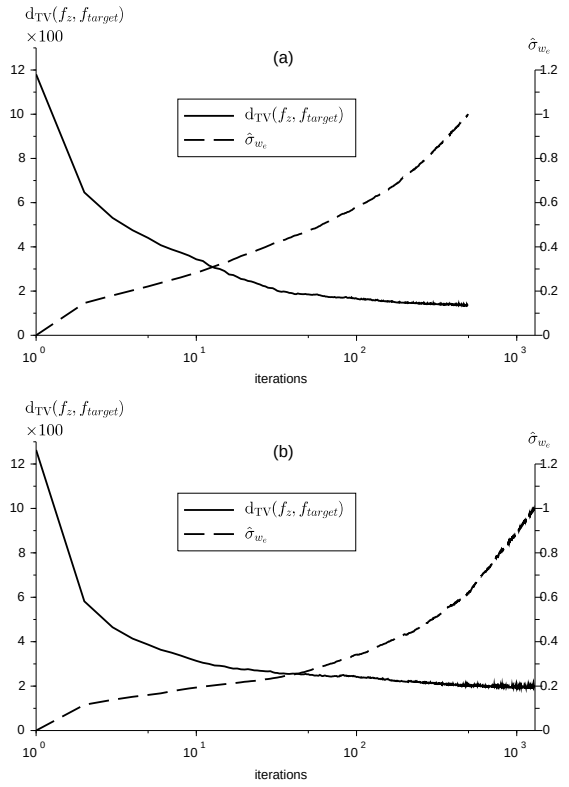

Fig. 14. *Histogram Local Reshaping* for a 3s speech sequence from a female (a) and a male (b) speakers, with target histogram defined by low-pass filtering of the original histogram, with cut-off frequency 1/32. Evolution of the Total Variation distance $\mathrm{d_{TV}}(f_z, f_{target})$ and of the estimated variance of w_e ($\hat{\sigma}_{w_e}$). According to the constraint (18), the algorithm stops when $\hat{\sigma}_{w_e}$ reaches 1.

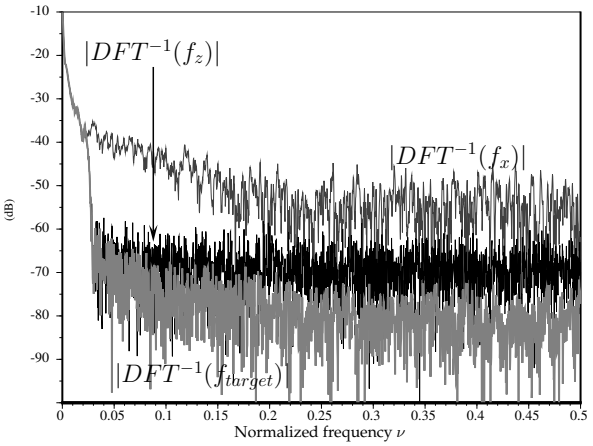

Fig. 15. After 100 iterations of the *HLR* algorithm, characteristic functions of the original signal ($|\mathrm{DFT}^{-1}(f_x)|$) and of the signal with low-pass-filtered histogram ($|\mathrm{DFT}^{-1}(f_z)|$), compared to the target characteristic function $|\mathrm{DFT}^{-1}(f_{target})|$.

TABLE III
PDF RECOVERY ERROR, DEFINED AS THE TOTAL VARIATION DISTANCE BETWEEN THE ORIGINAL AND THE RECOVERED PDF.

Signal	Recovery error	
	Female speaker	Male speaker
Original x	1.3×10^{-1}	1.3×10^{-1}
HLR-modified z	1.9×10^{-2}	2.7×10^{-2}
Transformed as in [10] y	2.4×10^{-3}	2.2×10^{-3}

parallel with the estimation of σ_{w_e}, for constant $\sigma_i = 0.25$, $MAX\_IT = \infty$ and $\Delta_d^{min} = 0$ (see Fig. 7). In each iteration, each second of signal needs 3s to be processed, with our Scilab implementation on a standard computer. After a strong decrease during the first iteration, the distance to the target histogram decreases more and more slowly, while the estimated σ_{w_e} reaches 1. The irregular evolution of $\mathrm{d_{TV}}(f_z, f_{target})$ for a large number of iterations is caused by the discontinuities smoothing procedure at the end of each iteration, which effect on $\mathrm{d_{TV}}$ may not be negligible anymore compared to the low decrease of $\mathrm{d_{TV}}$ during the iteration. If we stop the algorithm after 100 iterations, the histogram is really low-passed, as illustrated by Fig. 15 for the female speaker (a similar result was obtained for the male speaker), whereas z remains perceptually identical to x. The Mean Opinion Score (MOS) of z compared to x is estimated by PESQ [19]. The estimated MOS is 4.4 for the female speaker (4.3 for the male speaker), which indicates an inaudible distortion. Comparatively, low-pass filtering the histogram as in [10] leads to a signal y of which histogram f_y is equal to f_{target}, but with an estimated MOS of 3.9 for the female speaker and 3.7 for the male speaker (slightly audible distortion).

Note that $\delta(i,.)$ here is too far from a Bernoulli process, but is stationary, and the conditions (21,22) for whiteness of w_e are fulfilled for most iterations (see Appendix C).

We sub-quantized with a factor $K = 16$ the original signal x and its transformed versions z and y into x_Q, z_Q, and y_Q, respectively. Then we retrieved each original PDF from that of the quantized signal, using the filter G (32). The results are summarized in Table III.

The proposed algorithm leads to a recovery error which is much better than without low-pass filtering the histogram, higher than that of [10], but with a guaranteed inaudible transformation noise. Hence, HLR is an efficient approach for inaudibly lowpass-filtering sound histograms, which makes possible to at last take advantage from the powerful Widrow quantization theorem for audio signals.

V. CONCLUSION

We have proposed a sound histogram reshaping method that makes any histogram closer to a target histogram, while controlling the inaudibility of the transformation. The principle is to add iteratively a low-power white noise to a flat-spectrum version of the signal, until the target distribution or the noise audibility is reached. This scheme was applied through two variants, *Histogram Global Reshaping* (to change the global shape of the histogram) and *Histogram Local Reshaping* (to locally "chisel" the histogram, keeping its global shape unchanged).

This algorithm for perceptually controlled reshaping of sound histograms opens new perspectives in various audio processing applications where the original signal is available and its distribution preferably fulfills specific requirements: band-limited characteristic function for histogram restoration [10]; Gaussianity for non-linear audio system identification [5];

sparsity for informed audio source separation [12]; specific distribution for optimal audio quantization [13].

Future works will focus on reducing the complexity factor due to the length of the coloring filter impulse response. The solution could be based on a perception-based filter-bank (auditory filters mimicking the auditory system [30], [31]) and white noise addition in each frequency channel, which would avoid the memory effect responsible for complexity.

Audio signals have a malleability bounded by the properties of human audition, which was extensively exploited by audio-coding [32], [33] and watermarking [34]. This perceptually controlled reshaping of sounds histogram is another way of exploring this malleability.

Appendix A
Computation of $\sigma^2_{w_e}$

Let

$$w_e(n) = \sum_{i=1}^{q} \delta(i,n)\delta w_e(i,n), \tag{34}$$

where $\forall\, i$,

$$\begin{cases} \delta w_e(i,n) \sim \mathcal{N}(0,\sigma_i^2) \text{ and all } \delta w_e(i,n) \text{ are independent,} \\ \delta(i,.) \text{ is a Bernoulli process of parameter } p_i. \end{cases}$$

A. Case $\forall\, i, \quad p_i = p$ and $\sigma_i = \sigma$

The probability density function of $w_e(n)$ is given by:

$$f\Big(w_e(n)\Big) = \sum_{k=1}^{q} \Pr\left(\sum_{i=1}^{q}\delta(i,n)=k\right) f\left(w_e(n)\Big|\sum_{i=1}^{q}\delta(i,n)=k\right) \tag{35}$$

where

$$\left(w_e(n)\Big|\sum_{i=1}^{q}\delta(i,n)=k\right) \sim \mathcal{N}(0,k\sigma^2) \tag{36}$$

Hence, $w_e(n)$ has zero mean and variance:

$$\sigma^2_{w_e} = \sum_{k=1}^{q} \Pr\left(\sum_{i=1}^{q}\delta(i,n)=k\right) k\sigma^2 \tag{37}$$

$$= \sigma^2 \mathrm{E}\left[\sum_{i=1}^{q}\delta(i,n)\right] \tag{38}$$

$$= pq\sigma^2 \tag{39}$$

B. General case

For $(i_1 \dots i_q) \in \{0;1\}^q$, let us define the event:

$$\Lambda(n, i_1 \dots i_q) = \bigcap_{k=1}^{q} \{\delta(k,n) = i_k\} \tag{40}$$

The probability density function of $w_e(n)$ is given by:

$$f\Big(w_e(n)\Big) = \sum_{(i_1 \dots i_q)\in\{0;1\}^q} \Pr\big(\Lambda(n,i_1\dots i_q)\big) f\big(w_e(n)|\Lambda(n,i_1\dots i_q)\big) \tag{41}$$

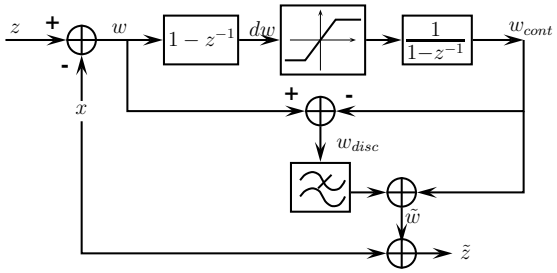

Fig. 16. Smoothing the discontinuities of the transformation noise w.

where

$$\big(w_e(n)\big|\Lambda(n,i_1\dots i_q)\big) \sim \mathcal{N}\left(0, \sum_{k=1}^{q} i_k\sigma_k^2\right) \tag{42}$$

Hence, $w_e(n)$ has zero mean and variance:

$$\sigma^2_{w_e} = \sum_{(i_1\dots i_q)\in\{0;1\}^q} \Pr\big(\Lambda(n,i_1\dots i_q)\big) \sum_{k=1}^{q} i_k\sigma_k^2 \tag{43}$$

Appendix B
Smoothing the discontinuities of the transformation noise w

The smoothing process is illustrated by Fig. 16. The general principle is to decompose w into two components, a continuous part w_{cont} and a piece-wise-constant signal w_{disc} (with jumps), and then to low-pass filter w_{disc}.

The transformation noise w is obtained by subtracting z to x and then differentiated. The difference signal dw is thresholded only at block transitions and then integrated, yielding w_{cont}. Subtracting the resulting signal to w yields a piece-wise-constant signal w_{disc}, containing the discontinuous component of w. Low-pass filtering w_{disc} and adding the result to w_{cont} provides a signal $\tilde{w}$, corresponding to w with smoothed block-transitions. Adding it to x provides the new z.

The thresholding is performed as follows. Let $dw(n) = w(n) - w(n-1)$. We define $dw_{\mathrm{med}}(n)$ as the median value of dw on the interval $n \pm M_1$, where M_1 is a small fixed integer. Let $\sigma_d^2(n)$ the estimated variance of $dw - dw_{\mathrm{med}}$ on the interval $n \pm M_2$, where M_2 is a fixed integer ($M_2 \gg M_1$). The bigger $|dw(n) - dw_{\mathrm{med}}(n)|$ is (relatively to $\sigma_d(n)$), the more probably a discontinuity occurred in w. Consequently, if $|dw(n) - dw_{\mathrm{med}}(n)| > 2\sigma_d(n)$, then $dw(n)$ is replaced by $dw_{\mathrm{med}}(n)$, otherwise it is let unchanged.

The low-pass filter has a triangular impulse response and is non-causal in such a way as to introduce no delay.

In the experiments described in Section IV, $M_1 = 2$, M_2 corresponds to 4ms and the low-pass filter has an impulse response of length 1ms.

Appendix C
Checking whether $\delta(i,.)$ is a Bernoulli process

For each iteration i, we want to check whether the decision $\delta(i,n)$ of adding $\delta_{w_e}(n)$ to $e_z(n)$ actually follows a Bernoulli process of parameter p_i. Denoting by T_i the random variable "number of zeros between two ones in $\delta(i,.)$", this means

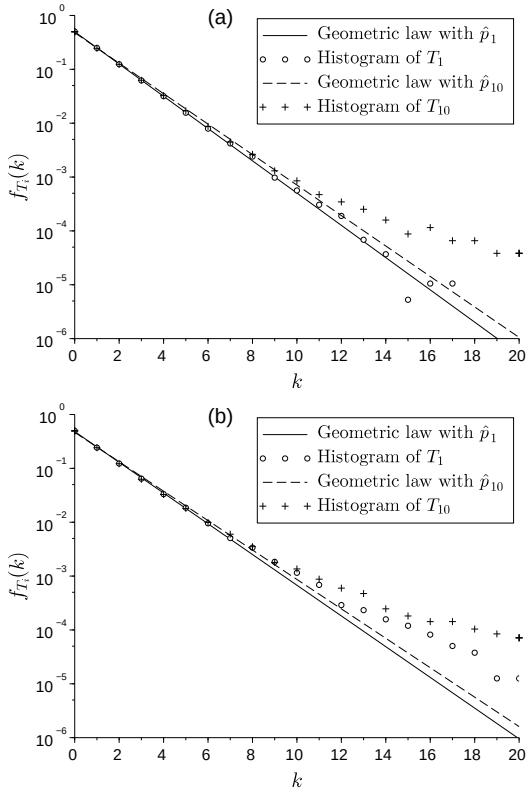

Fig. 17. How T_i fits a geometric law of parameter $\hat{p}_i$, for the first and last iterations, in the bass (a) and piano (b) experiments.

checking whether the $T_i(j)$ are independent and match a geometric law of parameter p_i.

We assessed the independence as follows [35]. First, a Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test checks the stationarity[11]. Then, if the process can be considered as stationary, the independence can be assessed through the auto-correlation coefficients $\rho(k)$: if the latter are inside the 95 % confidence interval around 0, the hypothesis of independence can be validated.

Assuming that T_i follows a geometric law, the estimation of the parameter p_i according to the maximum likelihood is given by $\hat{p}_i = 1/(1+\overline{T_i})$, where $\overline{T_i}$ is the empirical mean of T_i. We assessed the goodness of fit of the geometric law of parameter $\hat{p}_i$ graphically and through the chi-squared test for fit. For each i, we computed the test statistics for 1000 draws of 1000 individuals of T_i and observed the percentage of statistics χ^2 above the 5 % critical value.

The results are summarized in Table IV and Fig. 17, for the experiments of subsection IV-A. The assumption of independence for T_i is valid for both bass and piano. As illustrated by Fig. 17, the histogram of T_i matches well a geometric law, except in the tail of the distribution for the piano (for probabilities lower than 10^{-3}, emphasized here by the log-scale), which leads to poor results in the chi-squared tests.

For the experiments of subsection IV-B, the stationarity of

[11]We used the "grocer" toolbox of Scilab:
https://atoms.scilab.org/toolboxes/grocer

TABLE IV
RESULTS OF STATISTICAL TESTS ON T_i FOR THE BASS AND PIANO EXPERIMENTS. FOR THE KPSS TEST, $A_{x\%}$ DENOTES ACCEPTATION AT $x\%$ RISK AND R MEANS REJECTION. FOR THE AUTO-CORRELATION TEST, WE COMPUTED THE 100 FIRST AUTO-CORRELATION COEFFICIENTS. Q_{95} DENOTES THE 95 % CONFIDENCE INTERVAL.

	bass			piano		
Iteration	KPSS test	$\sharp\{\rho(k)$ out of $Q_{95}\}$	% of $\chi^2 > Q_{95}$	KPSS test	$\sharp\{\rho(k)$ out of $Q_{95}\}$	% of $\chi^2 > Q_{95}$
1	$A_{10\%}$	9	9	$A_{10\%}$	14	22
2	$A_{10\%}$	10	10	$A_{10\%}$	11	26
3	$A_{5\%}$	4	9	$A_{10\%}$	14	33
4	$A_{10\%}$	8	11	$A_{10\%}$	8	38
5	$A_{10\%}$	4	10	$A_{5\%}$	9	38
6	$A_{10\%}$	8	12	R	3	43
7	$A_{10\%}$	2	16	$A_{10\%}$	10	48
8	$A_{10\%}$	4	17	$A_{5\%}$	9	47
9	$A_{10\%}$	1	20	$A_{1\%}$	6	49
10	$A_{10\%}$	7	23	$A_{1\%}$	8	49

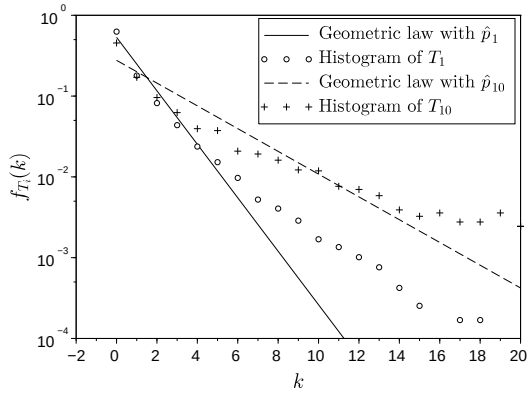

Fig. 18. How T_i fits a geometric law of parameter $\hat{p}_i$ in the female speaker experiment (first and 10^{th} iterations).

T_i is rejected for $i = 1$ (p_1 decreases during the iteration), and accepted at 1 to 10% risk for $i \geq 2$. For both speakers and all iterations, the majority of the auto-correlation coefficients are out of the 95% confidence interval around 0. Hence, the $T_i(j)$ cannot be considered as independent. Moreover, as illustrated by Fig. 18, the histogram of T_i does not match a geometric law: there are too much zeros and high values, *i.e.* $\delta(i,.)$ contains too long sequences of zeros or ones. Consequently, $\delta(i,.)$ here cannot be considered as a Bernoulli process.

For each of the 100 iterations, we computed the empirical values of $\mathrm{E}[\delta w_e(i,n) \,|\, \delta(i,n)=1]$ and $\mathrm{E}[\delta w_e(i,n)\delta w_e(i,n+k) \,|\, \delta(i,n)\delta(i,n+k) = 1]$ for k = 1 to 100 (see conditions (21,22)), and we counted how many of them are out of the 95 % confidence interval around zero. The results reported in Table V indicate that the conditions (21,22) are fulfilled or almost fulfilled for most iterations.

ACKNOWLEDGMENT

The authors thank the reviewers for their comments and suggestions, which were helpful to clarify and enrich this

TABLE V
NUMBER OF ITERATIONS (AMONG 100) HAVING EMPIRICAL $\mathrm{E}[\delta w_e(i,n) \mid \delta(i,n)=1]$ AND MORE THAN 5 OR 10 (AMONG 100) EMPIRICAL $\mathrm{E}[\delta w_e(i,n)\delta w_e(i,n+k) \mid \delta(i,n)\delta(i,n+k)=1]$ OUT OF THE 95 % CONFIDENCE INTERVAL AROUND ZERO.

	Number of iterations having		
	$\mathrm{E}[\delta w_e(i,n) \mid \delta(i,n)=1]$	more than 5 $\mathrm{E}[\delta w_e(i,n)\delta w_e(i,n+k) \mid \delta(i,n)\delta(i,n+k)=1]$	more than 10
	out of the 95 % confidence interval around 0		
female	9	31	0
male	1	45	0

article.

REFERENCES

[1] R. C. Gonzalez and R. E. Woods, Eds., *Digital Image Processing*, 3rd ed. Prentice Hall, 2008.

[2] S. A. White, "Restoration of nonlinearly distorted audio by histogram equalization," *J. Audio Eng. Soc*, vol. 30, no. 11, pp. 828–832, 1982.

[3] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 845–854, May 2006.

[4] R. Al-Wakeel, M. Shoman, M. Aboul-Ela, and S. Abdou, "Stereo-based histogram equalization for robust speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 15, june 2015.

[5] I. Mezghani-Marrakchi, G. Mahé, S. Djaziri-Larbi, M. Jaïdane, and M. Turki-Hadj Alouane, "Nonlinear audio systems identification through audio input Gaussianization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 41–53, january 2014.

[6] P. Balazs, B. Laback, G. Eckel, and W. A. Deutsch, "Time-frequency sparsity by removing perceptually irrelevant components using a simple model of simultaneous masking," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 18, no. 1, pp. 34–49, Jan. 2010.

[7] J. Pinel and L. Girin, "'Sparsification' of audio signals using the MDCT/IntMDCT and a psychoacoustic model – application to informed audio source separation," in *Proc. of the 42nd Audio Engineering Society Conference: Semantic Audio*, Ilmenau, Germany, 2011.

[8] G. Mahé, E. Nadalin, R. Suyama, and J. Romano, "Perceptually controlled doping for audio source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, p. 27, 2014.

[9] G. Chardon, T. Necciari, and P. Balazs, "Perceptual matching pursuit with Gabor dictionaries and time-frequency masking," in *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2014*, 2014.

[10] H. Halalchi, G. Mahé, and M. Jaïdane, "Revisiting quantization theorem through audiowatermarking," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, April 2009, pp. 3361–3364.

[11] B. Widrow, "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory," *IRE Transactions on Circuit Theory*, vol. 3, no. 4, pp. 266–276, December 1956.

[12] P. Comon and P. Jutten, *Handbook of Blind Source Separation*, P. Comon and P. Jutten, Eds. Academic Press, 2010.

[13] A. Gersho, "Principles of quantization," *Circuits and Systems, IEEE Transactions on*, vol. 25, no. 7, pp. 427–436, 1978.

[14] E. Zwicker and R. Feldtkeller, *The ear as a communication receiver*. New York: Acoustical Society of America, 1999.

[15] *Norm 11172-3: Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s – Part 3: Audio*, ISO/IEC, 1993.

[16] C. Baras, N. Moreau, and P. Dymarski, "Controlling the inaudibility and maximizing the robustness in an audio annotation watermarking system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1772–1782, Sept 2006.

[17] R. A. Roberts and C. T. Mullis, *Digital signal processing*. Addison-Wesley, 1987.

[18] ITU-R, "Recommendation BS.1387: Method for objective measurement of perceived audio quality," 1998.

[19] ITU-T, "Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," 2001.

[20] E. Vincent, S. Araki, F. J. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, pp. 1928–1936, 2012.

[21] A. Liutkus, R. Badeau, and G. Richard, "Gaussian processes for underdetermined source separation," *IEEE Transactions on Signal Processing*, vol. 59, pp. 3155–3167, 2011.

[22] M. K. Varanasi and B. Aazhang, "Parametric generalized Gaussian density estimation," *The Journal of the Acoustical Society of America*, vol. 86, no. 4, pp. 1404–1415, October 1989.

[23] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 107–115, May 2014.

[24] A. Liutkus, S. Gorlow, N. Sturmel, S. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed source separation: a comparative study," in *Proceedings European Signal Processing Conference (EUSIPCO 2012)*, Aug. 2012.

[25] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[27] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[28] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[29] B. Paillard, P. Mabilleau, S. Morissette, and J. Soumagne, "Perceval: perceptual evaluation of the quality of audio signals," *Journal of the Audio Engineering Society*, vol. 40, no. 1-2, pp. 21–31, 1992.

[30] S. Strahl and A. Mertins, "Analysis and design of gammatone signal models," *The Journal of the Acoustical Society of America*, vol. 126, pp. 2379–89, 2009.

[31] T. Necciari, N. Holighaus, P. Balazs, Z. Průša, P. Majdak, and O. Derrien, "Audlet filter banks: A versatile analysis/synthesis framework using auditory frequency scales," *Applied Sciences*, vol. 8, p. 96, Jan. 2018.

[32] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. John Wiley & Sons, Inc., 2005.

[33] N. Moreau, *Outils pour la compression des signaux - Applications aux signaux audio*. Hermès-Lavoisier, 2009.

[34] G. Hua, J. Huang, Y. Q. Shi, J. Goh, and V. L. Thing, "Twenty years of digital audio watermarking - a comprehensive review," *Signal Processing*, vol. 128, pp. 222 – 242, 2016.

[35] J.-Y. Le Boudec, *Performance Evaluation of Computer and Communication Systems*. EPFL Press, Lausanne, Switzerland, 2010.