Evaluation of short-time spectral attenuation techniques for the restoration of musical recordings

Olivier Cappé and Jean Laroche

TELECOM Paris, Département SIGNAL
46 Rue Barrault, 75634 Paris Cedex 13, France

Abstract

This paper deals with the application of short-time spectral attenuation techniques to the restoration of musical recordings degraded by background noise. Signal distortions induced by the restoration process are evaluated analytically, and their audibility is assessed on the basis of objective criteria. The results obtained highlight the influence of adjustable parameters (e.g., short-time frame duration or noise over-estimation) on the quality of the restoration.

I Introduction

Restoration of degraded audio recordings with the help of digital signal processing techniques is a field that has received a growing attention during the past years. In this paper, we will focus on a specific degradation of old recordings: background noise. For noise reduction in musical recordings, the techniques usually applied are based on *short-time spectral attenuation* (STSA). The general principle of these techniques consists in calculating a short-time transform of the noisy signal, and attenuating the values that are strongly corrupted by noise [1], [2], [3].

These techniques were originally introduced in the field of speech enhancement [1] and have been more recently applied to musical recordings [4], [5], [6]. As of today, most available results concerning the evaluation of such techniques have been obtained in the context of speech enhancement and consequently mainly concern speech intelligibility [7], [1]. Another aspect that is also taken into account is the quality of the enhanced speech, generally evaluated by listening tests [7], [2].

For the restoration of musical recordings the above criteria are not suitable: in particular, the expectation for quality is certainly higher for musical recordings than for speech. Furthermore, listening tests are made complex by the large variety of musical recordings (voice, orchestra and so on).

Previous work on the evaluation of restoration techniques include [4] and [3]. In both papers, the restoration technique is tested on a very simple signal consisting of a sinusoid embedded in noise. Two main conclusions are reached:

- Enhancement possibilities improve with the frequency selectivity of the short-time transform [4].
- The signal obtained at the output of the system is composed of the sinusoid surrounded by a narrow-band noise component [4], or equivalently, modulated by a noise [3]. However these two papers give very different results concerning the audibility of the phenomenon.

The purpose of this article is to test and extend these results on the basis of objective quality assessments. The approach followed involves two steps: first, we characterize analytically the undesirable alterations of the signal caused by the restoration process; second, we evaluate the audibility of the alterations by use of standard psychoacoustics results concerning simultaneous frequency masking. As we feel that discussions concerning the foundations and the limits of the auditory models used here would fall beyond the scope of the paper, we suggest that readers interested by these aspects refer to the classic textbooks [8], [9] and [10].

Note that in this paper the emphasis is put on signal distortion rather than noise reduction, since in practice signal distortion is the main limiting factor to the application of restoration techniques to musical signals. For results assessing the noise reduction obtained with such techniques, see [3] and [11].

The paper will be organized the following way: Section II presents simplifying hypotheses regarding both the test signals and the restoration techniques. Section III investigates the alterations caused by the restoration process itself, regardless of the random nature of the noise. Signal distortions caused by the presence of random noise are taken up in section IV.

II Preliminaries and hypotheses

II.A Short-time transform

In most of the systems mentioned above, the short-time transform used is the Short-Time Fourier Transform (STFT) [1], [7], [5]. In some systems [4], [3], a uniform filter bank is used which could also be implemented by use of the short-time Fourier transform [12]. In this paper we will consider the case of uniform short-time transforms such as the STFT. The following notations will be used:

- Short-Time Fourier Transform Analysis,

$$X(p, \phi_k) = \sum_{n=-\infty}^{+\infty} h(pR - n)x(n)e^{-\mathrm{j}2\pi\phi_k n} \tag{1}$$

- Short-Time Fourier Transform Synthesis,

$$y(n) = \sum_{p=-\infty}^{+\infty} f(n - pR)\frac{1}{N}\sum_{k=0}^{N-1} Y(p, \phi_k)e^{\mathrm{j}2\pi\phi_k n} \tag{2}$$

which corresponds to the 'fixed time-reference' definition used in [12]. $x(n)$ is the signal being analyzed, $h(n)$ and $f(n)$ are the so-called analysis and synthesis windows respectively. N is the size of the discrete Fourier transform, R is the hop size (in samples) between successive frames, $\phi_k = k/N$ is the k^{th} discrete normalized frequency. Thus $X(p, \phi_k)$ is the STFT at time index p and frequency ϕ_k. $Y(p, \phi_k)$ is the modified STFT, and $y(n)$ is the processed signal. By convention, $f(n)$ is zero for $n > 0$ and $h(n)$ is zero for $n < 0$.

In this paper, we will suppose that $h(n)$ and $f(n)$ are of length N as is usually the case in practice. Moreover, we will always use a hop factor R of 1 in order to simplify the calculations.

II.B Short-time spectral attenuation

In the restoration techniques mentioned above, the noise reduction is achieved by applying a real positive gain $G(p, \phi_k) < 1$ to the STFT of the noisy signal: $Y(p, \phi_k) = G(p, \phi_k) \times X(p, \phi_k)$. The gain is determined for each frame index p and each frequency ϕ_k by the so-called 'noise suppression rule' [3] as detailed below.

In most noise suppression rules, the gain $G(p, \phi_k)$ can be expressed as a function of the following quantity

$$Q(p, \phi_k) = \frac{|X(p, \phi_k)|^2}{L_d(\phi_k)} \tag{3}$$

in which $L_d(\phi_k)$ is the estimated noise level at frequency ϕ_k. $Q(p, \phi_k)$ will be called the 'relative signal level'. For example, the power subtraction rule [1], [2] is defined by

$$G(p, \phi_k) = \sqrt{1 - \frac{1}{Q(p, \phi_k)}} \tag{4}$$

and the so-called Wiener subtraction rule [2] is defined by

$$G(p, \phi_k) = 1 - \frac{1}{Q(p, \phi_k)} \tag{5}$$

In both cases, values of Q lower than 1 are artificially set to 1. The two preceding suppression rules are represented in Fig. 1. A modification of the noise suppression rule often used in practice consists in overestimating the noise level [1] [5]: $L_d(\phi_k)$ is replaced by $\alpha L_d(\phi_k)$ with $\alpha > 1$.

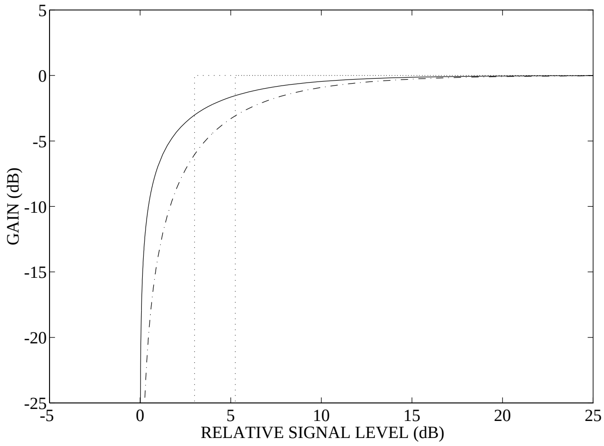

Figure 1: *Gain versus relative signal level; solid line: Power subtraction; dash-dot line: Wiener; dotted lines: ideal models corresponding to both suppression rules.*

In the following, in order to maintain the generality of the results and to simplify the calculations, the various noise suppression rules will be approximated by an 'ideal suppression rule' in which the gain is

$$\begin{cases} 0 & \text{when} \quad Q(p, \phi_k) < Q_c \\ 1 & \text{when} \quad Q(p, \phi_k) \geq Q_c \end{cases}$$

where Q_c is the relative cutoff level. Q_c is defined as that level $Q(p, \phi_k)$ for which the gain $G(p, \phi_k)$ is -3dB. The ideal suppression rules corresponding to the two preceding noise suppression rules are represented in dotted line in Fig. 1. $Q_c = 3$dB for the power subtraction rule and $Q_c = 5$dB for the Wiener suppression rule. It is easily verified that overestimating the noise by a factor α amounts to multiplying Q_c by α. Unless otherwise specified, the noise suppression rule will always be the ideal suppression rule.

II.C Hypotheses

In the following, the noise $d(n)$ will be considered stationary, additive and uncorrelated with the noise-free signal $s(n)$. Most of the time, the signal $s(n)$ considered in this paper will be composed of sinusoids with amplitudes and frequencies assumed constant within the analysis frame. This kind of signal makes it possible to highlight the limits of the restoration techniques considered here. Furthermore, for steady musical signals and standard frame durations (30 to 40 ms), this simple sinusoidal model often proves realistic [13], [14], [15].

Clearly, rapid musical transients (notes onsets, percussions) do not fall into the previous model. In this paper, to investigate the behavior of restoration techniques in such cases, we will use the example of a sinusoidal component with an abrupt onset. Although sinusoidal onsets cannot be considered an accurate model of musical transients, they provide an interesting insight into the limits of restoration techniques in the presence of such transients.

III Undesirable alterations of the signal resulting from the spectral attenuation

In this section, we investigate the effect of the restoration on the signal to be enhanced: according to the noise level, the spectral attenuation applied to the short-time spectra can alter the characteristics of the musical signal (i.e., the signal $s(n)$ above). Both stationary and transient signals are considered below.

The gain $G(p, \phi_k)$ is not a deterministic quantity as it is a function of the short-time power spectrum of the noisy signal. In the following sections, we will investigate the case of a deterministic gain equal to the mean value of $G(p, \phi_k)$. The consequences of the random nature of $G(p, \phi_k)$ will be taken up later.

III.A Timbre modification of stationary signals

We first consider a steady sinusoidal signal, and determine the power below which the sinusoid is annihilated by the restoration. This minimum power is then compared to auditory masking thresholds in order to determine whether the cancelled sinusoid was audible in the original noisy signal.

The signal $x(n)$ is assumed to be composed of a sinusoid of frequency Φ and amplitude A embedded in a noise $d(n)$ with power spectral density $P_d(\phi)$. We first calculate the mean value of the relative signal level Q at frequency Φ. Since the noise is uncorrelated with the sinusoid, we have

$$\begin{aligned} \mathrm{E}\left\{|X(p,\phi_k)|^2\right\} &= |S(p,\phi_k)|^2 + \mathrm{E}\left\{|D(p,\phi_k)|^2\right\} \\ &= |S(p,\phi_k)|^2 + L_d(\phi_k) \end{aligned} \tag{6}$$

in which $S(p, \phi_k)$ is the STFT of the sinusoid, and $\mathrm{E}\{|D(p, \phi_k)|^2\}$ is the mean value of the short-time power spectrum of the noise which we denoted $L_d(\phi_k)$. Note that since the noise is stationary, $\mathrm{E}\{|D(p, \phi_k)|^2\}$ is independent of time index p. If frequency Φ is larger than the width of the main lobe of the analysis window's Fourier transform, it is easily shown [16] that

$$|S(p,\Phi)| = \frac{A}{2} H(0) \tag{7}$$

where $H(0)$ is the Fourier transform of the analysis window at frequency 0:

$$H(0) = \sum_n h(n)$$

When the noise is white with variance σ^2, it is well known [16] that

$$L_d(\phi_k) = \mathrm{E}\left\{|D(p,\phi)|^2\right\} = \sigma^2 \left(\sum_n h(n)^2\right)$$

For large values of N, this result can be extended to noises with arbitrary power spectrum densities, under non restrictive hypotheses [17]:

$$L_d(\phi_k) = P_d(\phi) \left(\sum_n h(n)^2\right) \tag{8}$$

The mean value of the relative signal level at frequency Φ is derived from Eqs. (3) and (6) as

$$\mathrm{E}\left\{Q(p,\Phi)\right\} = 1 + \frac{|S(p,\Phi)|^2}{L_d(\phi_k)} \tag{9}$$

Inserting Eqs. (7) and (8) into the preceding equation gives

$$\mathrm{E}\left\{Q(p,\Phi)\right\} = 1 + \frac{A^2 \left[\sum_n h(n)\right]^2}{4P_d(\Phi)\left[\sum_n h(n)^2\right]} \tag{10}$$

Defining the equivalent noise bandwidth as in [16] by

$$\Delta_h = \frac{N\left[\sum_n h(n)^2\right]}{\left[\sum_n h(n)\right]^2}$$

Eq. (10) can be rewritten as

$$\mathrm{E}\left\{Q(p,\Phi)\right\} = 1 + \frac{A^2}{4} \times \left(P_d(\Phi) \times \frac{\Delta_h}{N}\right)^{-1} \tag{11}$$

This last equation is interpreted as follows. $A^2/4$ represents half the power of the sinusoid. Δ_h/N is the bandwidth (expressed in normalized frequency) of the rectangular bandpass filter equivalent to the analysis window $h(n)$ [16]. Therefore

$$P_d(\Phi) \times \Delta_h/N$$

represents the power of the noise in the STFT band centered around frequency Φ. Thus $\mathrm{E}\{Q(p,\Phi)\}$ can be interpreted as 1 plus the signal-to-noise ratio in the STFT band centered around Φ. Note that frequency Φ has been supposed to lie on a bin of the discrete Fourier transform. When Φ no longer lies on a STFT frequency bin and for standard analysis windows (Hanning, Hamming), the maximum deviation from the preceding result is -2dB [16].

The preceding result shows that multiplying the frame duration by a factor K raises the average relative signal level by the same factor. Note that it is the duration of the frame in seconds (not in samples) that needs to be taken into account.

During the restoration process, the sinusoid is cancelled if the signal level lies below the cutoff level Q_c. We define the restoration limit as the corresponding minimum signal power:

$$\frac{A^2}{4} = (Q_c - 1) \times \left(P_d(\Phi) \times \frac{\Delta_h}{N}\right) \tag{12}$$

We now consider a sinusoid whose amplitude is given by the preceding limit: it corresponds to the component with the smallest amplitude that is preserved in the restored signal. The question remains as to whether this component was audible in the original noisy signal. The answer to this question is found in studies pertaining to the masking of a pure tone by noise. Under the hypothesis that the power spectral density of the noise is constant around the sinusoidal frequency Φ, a classical result of psycoacoustics experiments [10], [8] states that the pure tone is masked if its power is less than that of the noise in the critical band centered around Φ. Because the width of the critical band varies with the center frequency [18], the masking threshold is a function of the pure tone frequency and is minimum for frequencies below 500Hz (for which the width of the critical bands is 100Hz).

From what precedes, to guarantee that the restoration process does not eliminate signal components audible in the noisy signal, it is sufficient to make sure that audible sinusoids with frequencies below 500Hz are unaltered. In other words, a component whose power is given by Eq. (12) should be below the minimum masking threshold:

$$\frac{A^2}{4} < P_d(\Phi)\frac{W_{cb}}{F_s} \tag{13}$$

in which W_{cb} designates the minimum width of the critical band and F_s is the sampling frequency, both expressed in Hz. Combining the preceding equation with Eqs. (13) and (12) we obtain

$$\frac{N}{F_s} > \frac{\Delta_h\,(Q_c - 1)}{W_{cb}} \tag{14}$$

This equation can be interpreted as follows: to avoid cancelling audible sinusoids during the restoration process, it is necessary to use an analysis window of sufficient duration. Standard orders of magnitude are 3-5dB for Q_c (see Fig. 1) and 1.5 for Δ_h [16] yielding a minimum window duration of about 40ms.

Note that the above minimum window duration is fairly long compared to standard durations used for speech enhancement (from 20 to 30ms). In addition, as the noise level $L_d(\phi_k)$ is often overestimated, the minimum window duration can be even higher.

Recall that the masking threshold increases with frequency: sinusoids with higher frequencies are more easily masked by the noise, and therefore one could tolerate that the restoration limit defined by Eq. (12) be an increasing function of frequency. This naturally leads to the idea of a short-time transform with a frequency-dependant spectral resolution. This has been exploited for example in [19] and [20]. Note however that such systems must exhibit a sufficient spectral resolution in the low frequency range: Eq. (14) can be rewritten in terms of spectral resolution as

$$\frac{F_s \Delta_h}{N} < \frac{W_{cb}}{Q_c - 1}$$

in which $F_s \Delta_h / N$ is the bandwidth of the equivalent rectangular analysis filter. With the same values as above, this maximum bandwidth is found to be about 40Hz in the low frequency range. This value is clearly below the bandwidth of the system proposed in [19].

According to Eq. (12), increasing the duration of the analysis window lowers the restoration limit, thus avoiding the cancellation of components with low signal levels. However, this is true only if the sinusoidal components are present during a time period longer than the duration of the analysis window. This is often the case for musical signals, for window durations considered above (less than 50ms). Finally, masking phenomena between sinusoidal components in the original signal have not been taken into account. For signals with a moderate noise level, such masking effects can contribute to conceal the disappearance of low amplitude components, thus loosening the constraint in Eq. (14).

III.B Spreading of transients

In this section, we study the effect of spectral attenuation on a complex sinusoidal component which appears at sample $n = 0$, with a level just above the restoration limit (Eq. (12)). We will make the following approximations:
1) The frequency Φ of the sinusoid lies on a bin of the discrete Fourier transform.
2) The gain $G(p, \phi_k)$ is null in all frames whose centers are located before sample $n = 0$. For frames whose centers are located right of $n = 0$, $G(p, \phi_k)$ is 1 for $\phi_k = \Phi$ and 0 elsewhere: Because the sinusoid lies just above the restoration limit, the only frequency bin at which the relative signal level is above Q_c is the one corresponding to Φ.

According to Eq. (25) (Appendix A), the modified signal $y(n)$ is

$$y(n) = \sum_{m-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(n-m) g_p(m) h(p-n+m) f(n-p) \tag{15}$$

with

$$\begin{cases} x(n) &= A \exp(\mathrm{j}2\pi\Phi n)\Pi(n) \\ g_p(m) &= \frac{1}{N}\exp(\mathrm{j}2\pi\Phi n)\Pi(p+\frac{N}{2}) \end{cases} \tag{16}$$

where $\Pi(n)$ is the step function ($\Pi(n) = 1$ if $n \geq 0$ and 0 elsewhere). The term $\Pi(n)$ in $x(n)$ indicates that the sinusoid starts at time $n = 0$. The term $\Pi(p + \frac{N}{2})$ in $g_p(m)$ comes from the fact that the gain is null for frames whose center is left of $p = 0$. $g_p(m)$ is the periodic impulse response corresponding to the frequency gain $G(p, \phi_k)$ defined by Eq. (23). Inserting Eq. (16) into Eq. (15) gives

$$\begin{aligned} y(n) &= A\exp(\mathrm{j}2\pi\Phi n) \\ &\times \frac{1}{N} \sum_{p=-\infty}^{+\infty} f(n-p)\Pi(p+\frac{N}{2}) \\ &\times \sum_{m-\infty}^{+\infty} h(p-n+m)\Pi(n-m) \end{aligned} \tag{17}$$

or

$$\begin{aligned} y(n) &= A\exp(\mathrm{j}2\pi\Phi n) \\ &\times \frac{1}{N} \sum_{p=-\infty}^{+\infty} f(n-p)\Pi(p+\frac{N}{2}) h * \Pi(p) \end{aligned} \tag{18}$$

Let us define $M(p) = \Pi(p + \frac{N}{2})h * \Pi(p)$. $M(p)$ is the truncated convolution of $h(p)$ and $\Pi(p)$. We now have

$$y(n) = A\exp(\mathrm{j}2\pi\Phi n)\left(\frac{1}{N}f * M(n)\right) \tag{19}$$

It is seen the the output signal is composed of a sinusoid with a time-envelope $\{f * M(n)\}/N$. It is easy to verify that $f * M(n)$ is null left of $n = -N/2$ and constant right of $n = +N$ (recall that $h(n)$ is zero outside of $[-(N-1), 0]$ and $f(n)$ is zero outside $[0, (N-1)]$). This shows that the abrupt transient occurring at $n = 0$ in the original signal is now spread between $n = -N/2$ and $n = +N$. Fig. 2 presents the aspect of the time-envelope for standard windows ($h(n)$ is a 128pt Hanning window and $f(n)$ is a 128pt rectangular window).

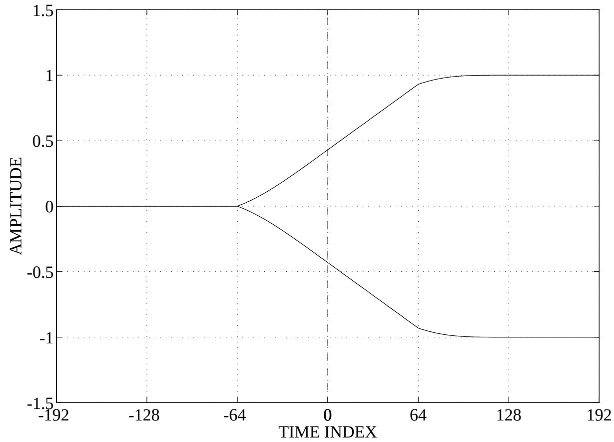

Figure 2: *Envelope of the restored signal for a component at the restoration limit. STFT parameters: $h(n)$ Hanning window, $f(n)$ rectangular window, with $N = 128$pts. Ideal suppression rule.*

Fig. 2 exemplifies the problems encountered with abrupt transients near the restoration limit. The visible transient spreading is due to the high selectivity of the spectral modification (only one frequency bin is unattenuated). The order of magnitude of the spread is $1.5N$ *for sinusoids near the restoration limit.*

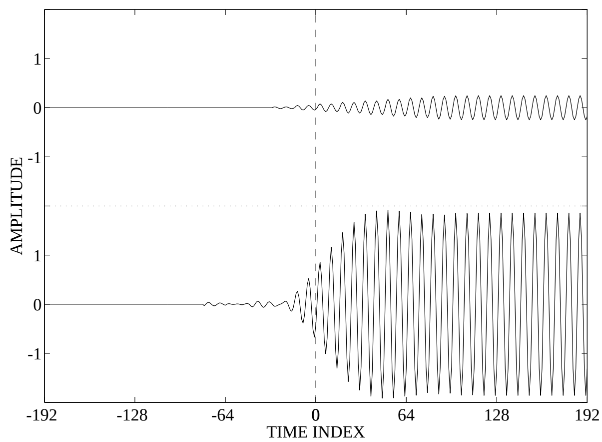

Figure 3: *Restored signals for two different levels: top, at the restoration limit; bottom, 10dB above the restoration limit. STFT parameters: $h(n)$ Hanning window, $f(n)$ rectangular window, with $N = 128$pts.*

When the relative level of the sinusoid increases, more frequency bins become unattenuated making the spectral modification less selective. As a consequence, the transient is spread over a smaller time interval. This is shown in Fig. 3 for two sinusoids with relative levels 3dB (upper curve) and 13dB (lower curve).

For the sake of clarity, the non-noisy signal $s(n)$ was processed in place of the noisy signal $x(n)$, while the noise level estimate $L_d(\phi_k)$ was kept unchanged : it is easily verified that with the power subtraction rule, processing the non-noisy signal is equivalent *in average* (and in average only) to processing the noisy signal with a 3dB (factor 2 on a linear scale) noise overestimation. This procedure highlights the spreading of the transient due to the spectral modification by eliminating the amplitude modulations due to the random nature of the noise (see section IV). With the suppression rule used here (power subtraction) the level of the component shown in the upper part of Fig. 3 corresponds to the restoration limit. In this case, one verifies that the transient spreading is similar to that displayed on Fig. 2. By contrast, the bottom signal in Fig. 3 exhibits a steeper transient. In all cases, the original transient is spread in a non-causal way, although mainly to the right of the onset time. Note that the shapes of the real signals shown on Fig. 3 differ slightly from the envelope of Fig. 2 because the behavior of the gain $G(p, \phi_k)$ is more complex than the simple jump model that we considered in Eq.(16).

When the window duration is increased, two opposite factors come into play: 1) The relative level of the sinusoidal component increases, which tends to minimize the amount of spreading (Eq. 11). 2) As was the case for the component near the restoration limit, the spread duration is quasi proportional to the window duration. Except for components close to the restoration limit, the second factor is prominent: Increasing the duration of the analysis window tends to increase the amount of transient spreading. This phenomenon imposes an upper limit on the window duration according to the nature of the signal. For example, for a singer with orchestra, time durations up to 80ms can be considered. By contrast, piano recordings require durations below 50ms.

An interesting point is that similar temporal artefacts occurring around transients have been reported to appear in wideband transform coders [21] or in systems for time-scale modification of speech based on the STFT [22]. These effects are all caused by modifications of the short-time transform of the signal and are thus closely dependent on the time resolution of the short-time transform [21]. However they differ by the exact nature of the modification brought to the short-time transform : for wideband coders, the pre-echoes observed correspond to a spreading of the quantification noise [21]; for time-scale modification systems, the temporal distortions are due to an incorrect STFT phase [22]; for music enhancement, the transient spreading appears to be mainly a consequence of the bandpass filtering effect of the modification brought to the STFT of the signal.

IV Influence of the random nature of noise

In section III-A, we determined the conditions under which a signal component is eliminated by the restoration process. We now study the case of a steady component that is above the restoration limit. By contrast with the preceding section, the random nature of the noise and of the gain $G(p, \phi_k)$ will be taken into account. The first subsection describes the spectral content of the denoised sinusoid. The second subsection studies the level of fluctuation of the gain $G(p, \phi_k)$ as a function of the relative signal level.

IV.A Perturbations caused by the noise remaining around components

Consider the case of a steady sinusoidal component of frequency Φ above the restoration limit. The gain $G(p, \phi_k)$ is then independent of time index p and is 0 except for few values around Φ where it is 1. When the spectral attenuation is time-independent and the hop factor R is 1, the STFT modification can be shown to be equivalent to a linear, time-invariant filter whose impulse response is given in Appendix A. As a result, the restored component is composed of a sinusoid surrounded by a band of noise centered at frequency Φ. As was pointed out in [4], the restoration gives satisfactory results only if this noise is inaudible, i.e. if it is masked by the sinusoid. To determine the audibility of the noise, we use the results in [23].

The first step consists in calculating the power spectral density of the noise remaining around Φ. This power spectral density depends on the number of frequency bins ϕ_k where the gain is 1 (unattenuated bins). Table 1 gives the maximum number of such bins for common analysis windows (Hanning, Hamming).

The power spectral density of the noise remaining around Φ is obtained the following way: first, the

Peak level Q	Maximum number of bins
$Q/Q_c < 6\text{dB}$	2
$6\text{dB} < Q/Q_c < 15\text{dB}$	3
$15\text{dB} < Q/Q_c < 30\text{dB}$	4

Table 1: *Maximum number of STFT bins above the cutoff level Q_c as a function of the sinusoid's relative level.*

transfer function equivalent to one unattenuated channel is calculated (Fourier transform of Eq. (30) in Appendix A). Then, according to the number of unattenuated channels (Table 1) the transfer functions of the channels around Φ are summed and multiplied by the original noise power spectral density at frequency Φ (the original noise is assumed to have a power spectral density constant over the unattenuated bins). The following figures (Figs. 4, 5 and 6) correspond to the case where $P_d(\Phi) = 30\text{dB}$.

Fig. 4 presents the power spectral density of the noise remaining around Φ for different values of the number of unattenuated bins and of the duration of the analysis frame. Note that as expected, longer frame durations correspond to narrower power spectral densities. Moreover, the width of the main lobe increases with the number of unattenuated bins.

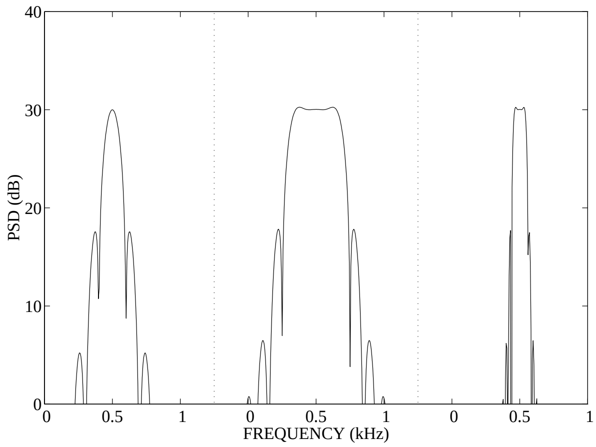

Figure 4: *Power spectral density of the remaining noise: left, frame duration of 10ms with 1 unattenuated bin; center, frame duration of 10ms with 4 unattenuated bins; right, frame duration of 40ms with 4 unattenuated bins. In all three cases the center frequency Φ is 500Hz. $h(n)$ is a Hanning window and $f(n)$ a rectangular window.*

The procedure described in [23] to determine the audibility of the noise goes as follows: for both the sinusoid and the noise, excitation patterns are calculated in a Bark frequency scale [18], [10]. A 'sensitivity function' is then applied to the excitation pattern of the sinusoid to obtain the masking threshold: the noise is inaudible if its excitation pattern is below the sinusoid's masking threshold for any given frequency.

Figs. 5 and 6 present the masking thresholds (dashed lines) and excitation patterns (solid lines) in four different cases: the left curves (respectively the right curves) correspond to a sinusoid of relative level 10dB (respectively 30dB). Fig. 5 corresponds to a frame duration of 10ms and Fig. 6 to a frame duration of 40ms. Clearly, for a frame duration of 10ms, the remaining noise is still audible even when the relative level is high (30dB). By contrast, for a frame duration of 40ms, the noise is audible for small relative levels (10dB) but masked for higher relative levels.

Moreover, the shapes of the noise's excitation patterns differ significantly for frame durations 40ms and 10ms. For a frame duration of 10ms, the excitation pattern of the noise is broader than masking threshold of the sinusoid. In Fig. 5 the width of the noise excitation pattern is much larger than a critical bandwidth (100Hz for a center frequency of 500Hz) and the remaining noise is distinctly heard as an additional narrow-band noise. By contrast, in Fig. 6 the excitation pattern of the noise has a similar shape as the masking

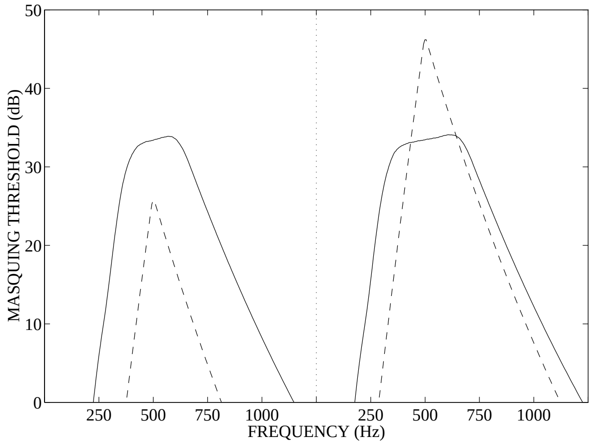

Figure 5: *Masking threshold of the pure tone of frequency* $\Phi = 500Hz$ *(dashed curve) compared to the excitation pattern of the remaining noise (solid curve) for a frame duration of 10ms: left, tone with a relative level of 10dB; right, tone with a relative level of 30dB.*

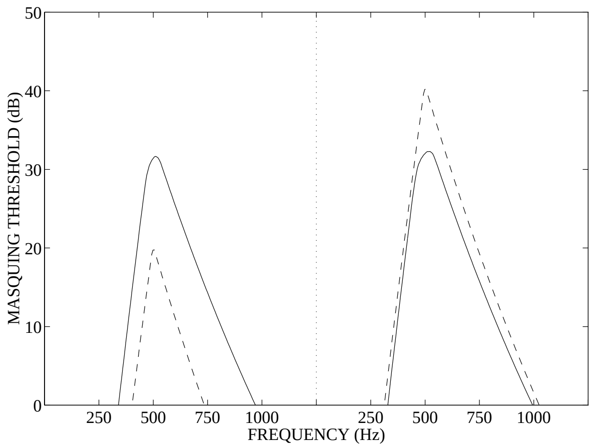

Figure 6: *Masking threshold of the pure tone of frequency* $\Phi = 500Hz$ *(dashed curve) compared to the excitation pattern of the remaining noise (solid curve) for a frame duration of 40ms: left, tone with a relative level of 10dB; right, tone with a relative level of 30dB.*

threshold of the sinusoid. When the remaining noise is audible, it is heard as an erratic fluctuation of the sinusoid level.

As the frequency of the sinusoid increases, the noise is more easily masked due to the widening of the critical bands. The examples presented above correspond to the worst case.

To summarize the preceding results, it should be emphasized that for window durations smaller than about 30ms, the noise remaining around sinusoidal components is likely to give birth to very undesirable audible effects even for sinusoids of high relative levels. The unpleasant nature of these alterations comes mainly from the fact that the narrow-band remaining noise depends on the signal component, which was not the case for the original wide-band noise. As a consequence, systems with an insufficient frequency resolution are prone to exhibit such artifacts as was found experimentally in [5]. For window durations above 40ms and relative levels below about 25dB, the remaining noise causes an audible modulation of the signal component.

In practice however, the audibility limits found above should be relaxed for musical signals, especially for very low noise levels: the time-varying nature of musical signals as well as the possible masking between signal components help conceal the undesirable effects observed with very simple signals.

IV.B Fluctuation of the component level

In the preceding sections, the relative signal level (and therefore the gain $G(p, \phi_k)$) were given deterministic values. This is not the case in practice: due to the noise in the signal, the spectral gain applied to a steady signal component undergoes random variations which then modulate the amplitude of the restored component. An example is given below. A sinusoid of fixed frequency and exponential decay initially 20dB above the noise level is restored using the power subtraction suppression rule (with a noise overestimation of $\alpha = 6$dB). Fig. 7 shows the gain at the sinusoidal frequency as a function of time. Note that the abscissa of Fig. 7 does not correspond to the time index p but to the *average value* of the relative signal level at time p. For high signal levels, the gain is close to 1 and fairly stable. As the relative signal level decreases, the gain undergoes more and more fluctuations, and finally exhibits an erratic behavior (including null values) for low relative signal levels. As a consequence, the level of the restored sinusoid undergoes fluctuations for relative signal levels between 10 to 20dB, and the sinusoid becomes intermittent for lower signal levels. When audible, this effect causes undesirable disturbances.

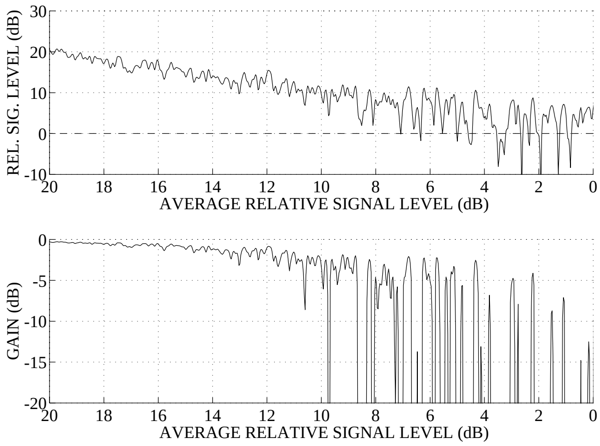

Figure 7: *Restoration of an exponentially decaying sinusoid: top, relative signal level measured at the frequency of the sinusoid versus its mean value; bottom, corresponding gain. STFT parameters: $h(n)$ Hanning, $f(n)$ rectangular, $N = 512$pts and $R = N/8$ (total number of time indexes displayed around 750). Suppression rule: power subtraction with 6 dB overestimation.*

To give an estimate of the range within which this fluctuation effect is observed, we need to estimate the probability density $f(Q)$ of the relative signal level. This calculation, described in appendix B, yields

$$f(Q) = e^{-[Q+(\bar{Q}-1)]} I_0 \left(2\sqrt{Q\,(\bar{Q}-1)} \right) \tag{20}$$

Where $I_o(x)$ denotes the modified Bessel function of order 0 and $\bar{Q}$ is the average relative level. The probability densities corresponding to various values of $\bar{Q}$ are illustrated in Fig. 8.

By numerically integrating the probability density defined by Eq. (20), the 99.9% confidence interval can be determined for each value of $\bar{Q}$. Recall that the confidence interval is the range within which Q lies with a probability of 0.999. The confidence intervals are depicted in Fig. 9 for various values of $\bar{Q}$.

Depending on the value of $\bar{Q}$ two different types of gain fluctuations are observed:

1) For values of $\bar{Q}$ below 10dB, there is a significant probability that Q lie below 3dB (the 5 first vertical solid lines in Fig. 9 extend below 3dB). In such cases, when the ideal suppression rule is used, the gain is null with the same probability. This can be observed in Fig. 7: when $\bar{Q}$ is lower than 10dB, the gain occasionally drops to null values. In practice, the restored signal exhibits intermittent components.

2) For values of $\bar{Q}$ above 10dB, there is a negligible probability that the gain become null and the component is always present in the restored signal. However, if the suppression rule used is not the ideal one, a fluctuation of the amplitude's component is still observed. For example, with the power subtraction rule, one can verify

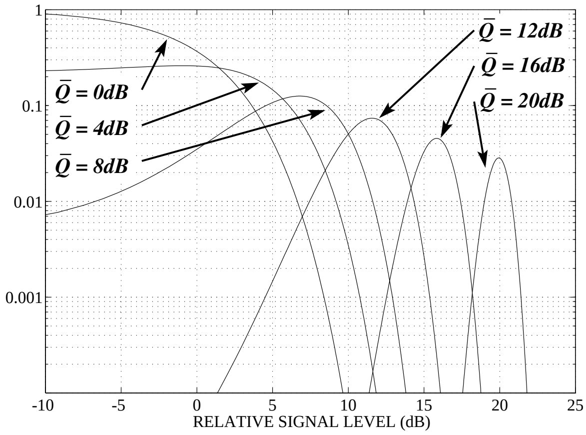

Figure 8: *Probability density of the relative signal level for different mean values $\bar{Q}$ (from left to right: 0, 4, 8, 12, 16 and 20dB).*

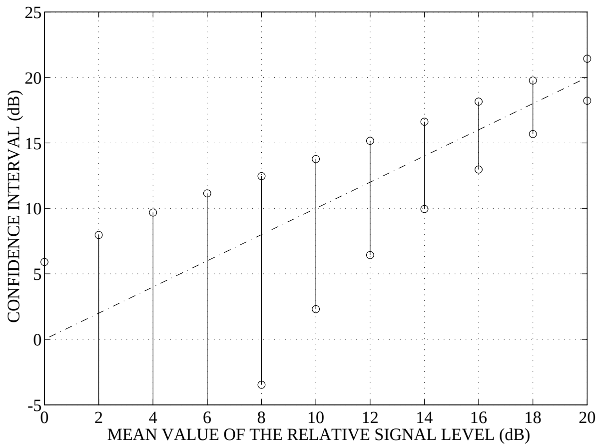

Figure 9: *99,9% confidence interval for the relative signal level versus its mean value. The dash-dot line indicates the mean value. For the first four intervals (mean values of 0, 2, 4 and 6dB) the lower bound of the confidence interval is not visible.*

that the fluctuation of the gain is less than 1dB (the order of magnitude of the smallest detectable change of amplitude [10]) when the relative level is above 7dB. Thus the gain fluctuation has no audible consequence when the lower limit of the confidence interval is above 7dB, or equivalently, when $\bar{Q} \geq 14$dB (Fig. 9). This limit is lower than that found in section IV-A: in practice, the fluctuation of the gain has a smaller impact on the restored signal than the remaining noise around sinusoidal components.

Finally, the analytical expression for the probability density of the relative signal level Eq. (20) can also be used to adjust the thresholds used in various ad hoc methods designed to eliminate erratic gain fluctuations [7], [5].

V Conclusion

In conclusion, the following important points can be underlined:

- The use of a simplified model of the restoration process has made it possible to bring to light and quantify the distortions caused by short-time attenuation techniques.
- Furthermore, the use of objective criteria regarding audibility limits has given a rational basis to

several results observed experimentally in previous papers. However, the perceptual evaluation of the distortions caused to transient signals could not be carried-out: In addition to the difficulty of obtaining an analytic description in that case, the question remains as to whether the test signal used (a sinusoid with an abrupt onset) is representative of musical transients.

- Our results give a quantitative evaluation of the well known compromise between small frame-duration and high frequency-resolution in the context of short-time spectral attenuation: short analysis-frames must be used to avoid spreading transients, and a high spectral resolution is needed to obtain a satisfying enhancement during quasi-stationary portions of the signal. A new important result is that when the noise level is significant, analysis-frames of too short durations (20ms or less) are bound to cause unacceptable distortions in the restored signal.

A consequence of the preceding points is that for certain types of musical signals, the above compromise cannot be satisfied: to reduce the loss of spectral components, frame durations as long as 40-50ms are needed, which can cause unpleasant audible distortions if the original recording features sharp transients. We strongly believe that improvements in this domain are made possible by the design of analysis/synthesis schemes that explicitly take into account both the transient and the quasi-stationary nature of musical signals. Examples of such schemes are the use of short-time transforms with a non-uniform frequency resolution [20] or the local variation of the frequency resolution of the short-time transform according to the characteristics of the processed signal [24], [25].

Acknowledgment

The authors would like to thank one of the anonymous reviewers for pointing out some important references.

A Constant modification of the short-time Fourier transform

We consider here the effect of a multiplicative modification of the short-time Fourier transform. We first examine the general case of a time-varying gain $G(p, \phi_k)$. We then restrict ourselves to the case of a time-invariant modification $G(\phi_k)$. The following derivations are inspired by [12].

The modified STFT is obtained as

$$Y(p, \phi_k) = X(p, \phi_k)G(p, \phi_k)$$

From Eq. (2), the modified signal is then

$$y(n) = \sum_{p=-\infty}^{+\infty} f(n-p) \frac{1}{N} \sum_{k=0}^{N-1} X(p, \phi_k)G(p, \phi_k)e^{\mathrm{j}2\pi\phi_k n} \tag{21}$$

Where de STFT of the noisy signal $X(p, \phi_k)$ is given by Eq. (1). Combining Eqs. (21) and (1) yields

$$\begin{aligned} y(n) &= \sum_{l=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(l)h(p-l)f(n-p) \\ &\quad \times \left\{ \frac{1}{N} \sum_{k=0}^{N-1} G(p, \phi_k)e^{\mathrm{j}2\pi\phi_k(n-l)} \right\} \end{aligned} \tag{22}$$

We define

$$g_p(m) = \frac{1}{N} \sum_{k=0}^{N-1} G(p, \phi_k)e^{\mathrm{j}2\pi\phi_k m} \tag{23}$$

$g_p(m)$ is the infinite impulse response obtained by inverse discrete Fourier transform of the spectral modification $G(p,\phi_k)$. Inserting the definition of $g_p(m)$ in Eq. (22) leads to

$$y(n) = \sum_{l=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(l)h(p-l)f(n-p)g_p(n-l) \tag{24}$$

Using the notation $m = n - l$, the previous equation can be rewritten as

$$y(n) = \sum_{m=-\infty}^{+\infty} \sum_{p=-\infty}^{+\infty} x(n-m)g_p(m)h(p-n+m)f(n-p) \tag{25}$$

If we now consider the case of a spectral modification constant over all short-time frames, the equivalent impulse response is given by Eq. (23) as:

$$g(m) = \frac{1}{N} \sum_{k=0}^{N-1} G(\phi_k)e^{\mathrm{j}2\pi\phi_k m}$$

where the time index p has been omitted. Eq. (25) can then be written as

$$y(n) = \sum_{m=-\infty}^{+\infty} x(n-m)g(m) \left(\sum_{p=-\infty}^{+\infty} h(p-n+m)f(n-p) \right) \tag{26}$$

The left hand term (in parenthesis) is recognized as the convolution of the two windows $h(n)$ and $f(n)$. Therefore the signal $y(n)$ obtained by modifying the STFT of $x(n)$ is

$$y(n) = \sum_{m=-\infty}^{+\infty} x(n-m) \left\{ g(m) \left[h * f(m) \right] \right\} \tag{27}$$

which indicates that the STFT modification amounts to linear filtering with the following impulse response

$$\tilde{g}(m) = g(m) \left[h * f(m) \right] \tag{28}$$

Note that when only one channel of frequency Φ is unattenuated, the inverse discrete Fourier transform of $G(\phi_k)$ is

$$g(m) = \frac{1}{N} e^{\mathrm{j}2\pi\Phi m} \tag{29}$$

and we have

$$\tilde{g}(m) = e^{\mathrm{j}2\pi\Phi m} \left\{ \frac{f * h(m)}{N} \right\} \tag{30}$$

B Calculation of the Probability Density of the relative level Q

Under the hypothesis of a stationary noise and provided N is large enough, the short-time Fourier transform of the noise alone at frequency ϕ_k and time p is a normally distributed complex variable [17] with variance $L_d(\phi_k)$.

For the sake of clarity, the frequency ϕ_k and the time index p will now be omitted. The STFT of the noisy signal, can be written as

$$X = S + D$$

where S is the STFT of the deterministic signal, and D the STFT of the noise. The relative signal level is defined as

$$Q = \frac{|S+D|^2}{L_d} \tag{31}$$

As was said before, the noise contribution D is a complex normally distributed variable: its real and imaginary parts (denoted D_r and D_i) are independent centered gaussian variables with variance $L_d/2$. Note that one can assume without loss of generality that the signal component S is real and positive as a rotation does not modify the value of Q. The relative signal level can thus be written

$$Q = \frac{(D_r + S)^2 + D_i^2}{L_d} \tag{32}$$

The joint distribution of D_r and D_i is given by

$$f(D_r, D_i) = \frac{1}{\pi L_d} e^{-\left[\frac{D_r^2 + D_i^2}{L_d}\right]}$$

To obtain the probability density of Q we use the classical change of variables [26] $(D_r, D_i) \mapsto (Q, \theta)$:

$$\begin{cases} D_r &= \sqrt{L_d}\sqrt{Q}\cos\theta - S \\ D_i &= \sqrt{L_d}\sqrt{Q}\sin\theta \end{cases}$$

The corresponding Jacobian is equal to $L_d/2$. The joint distribution of Q and θ is therefore

$$f(Q, \theta) = \frac{1}{2\pi} e^{-\left[L_d Q - 2\sqrt{L_d} S \sqrt{Q}\cos\theta + S^2\right]/L_d} \tag{33}$$

We can express the preceding equation in terms of $\bar{Q}$ the mean value of Q, obtained from Eq. (9) as $\bar{Q} = S^2/L_d + 1$:

$$f(Q, \theta) = \frac{1}{2\pi} e^{-\left[Q - 2\sqrt{Q(\bar{Q} - 1)}\cos\theta + (\bar{Q} - 1)\right]}$$

The probability density of the relative signal level is obtained by integrating over all the possible values of θ:

$$f(Q) = e^{-[Q + (\bar{Q} - 1)]} \times \frac{1}{2\pi} \int_0^{2\pi} e^{2\sqrt{Q(\bar{Q} - 1)}\cos\theta} \, d\theta \tag{34}$$

The right hand term is recognized as the modified Bessel function of order 0 [27], [26]. Finally, the probability density of the relative signal level Q is

$$f(Q) = e^{-[Q + (\bar{Q} - 1)]} I_0\left(2\sqrt{Q(\bar{Q} - 1)}\right) \tag{35}$$

where $I_o(x)$ denotes the modified Bessel function of order 0.

References

[1] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, December 1979.

[2] R. J. Mc Aulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(2):137–145, April 1980.

[3] P. Vary. Noise suppression by spectral magnitude estimation – Mechanism and theoretical limits. *Signal Processing*, 8(4):387–400, 1985.

[4] R. Lagadec and D. Pelloni. Signal enhancement via digital signal processing. In *Preprints of the AES 74th Convention*, New York, 1983.

[5] J. A. Moorer and M. Berger. Linear-phase bandsplitting: Theory and applications. *J. Audio Eng. Soc.*, 34(3):143–152, 1986.

[6] S. Vaseghi and R. Frayling-Cork. Restoration of old gramophone recordings. *J. Audio Eng. Soc.*, 40(10):791–801, 1992.

[7] S. F. Boll. Suppression of acoustic noise in speech using spectral substraction. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(2):113–120, 1979.

[8] B. C. J. Moore. *An introduction to the psychology of hearing.* Academic Press, second edition, 1982.

[9] B. Scharf. Critical bands. In J. V. tolbias, editor, *Foundations of modern auditory theory.* Academic Press, New York, 1970.

[10] E. Zwicker and R. Feldtkeller. *Das ohr als nachrichtenempfänger* (french translation). Masson, 1981.

[11] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. To appear in IEEE Trans. Speech and Audio Processing, April 1994.

[12] R. E. Crochiere and L. R. Rabiner. *Multirate digital signal processing.* Prentice-Hall, 1983.

[13] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music J.*, 14(4):12–24, Winter 1990.

[14] E. B. George and M. J. T. Smith. Analysis-by-synthesis/Overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *J. Audio Eng. Soc.*, 40(6):497–516, 1992.

[15] Diana Deutsch, editor. *The psychology of music.* AP series in cognition and perception. Academic Press, 1982.

[16] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proc. IEEE*, 66(1):51–83, 1978.

[17] D. R. Brillinger. *Time Series Data Analysis and Theory.* Holden-Day, expanded edition, 1981.

[18] E. Zwicker and E. Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *J. Acoust. Soc. Am.*, 68(5):1523–1525, 1980.

[19] T. L. Petersen and S. F. Boll. Acoustic noise suppression in the context of a perceptual model. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1086–1088, 1981.

[20] J. C. Valière, S. Montresor, and J. F. Allard. Présentation d'une méthode de suppression des bruits de surface sur les anciens enregistrements de musique. In *Colloque de physique C2, supplément au n 2, tome 51*, pages 761–764. Premier Congrès Français d'Acoustique, Février 1990.

[21] J. D. Johnston and K. Brandenburg. Wideband coding–Perceptual considerations for speech and music. In S. Furui and M. M. Sondhi, editors, *Advances in speech signal processing.* Marcel Dekker, New York, 1992.

[22] D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust., Speech, Signal Processing*, 32(2):236–242, April 1984.

[23] M. R. Schroeder, B. S. Atal, and J. L. Hall. Optimising digital speech coders by exploiting masking properties of the human ear. *J. Acoust. Soc. Am.*, 66(6):1647–1652, 1979.

[24] O. Cappé. Enhancement of musical signals degraded by background noise, using long-term behavior of the short-term spectral components. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages I–217–I–220, 1993.

[25] O. Cappé. *Techniques de réduction de bruit pour la restauration d'enregistrements musicaux (TELECOM Paris 93 E 019).* PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1993.

[26] A. Papoulis. *Probability, random variables, and stochastic processes.* McGraw-Hill, New York, 3rd edition, 1991.

[27] M. R. Spiegel. *Mathematical Handbook of Formulas and Tables.* Schaum's outline series. McGraw-Hill, 1968.