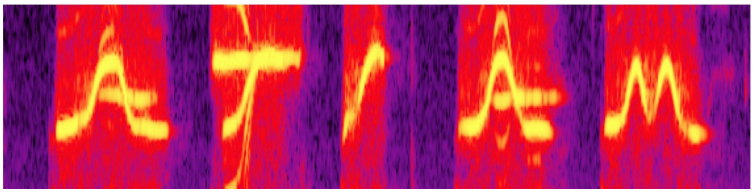
Sound Texture Synthesis Informed by Statistics

Hugo Saulnier

March-July 2013
IRCAM, Analysis-Synthesis Team,
under the direction of Axel Röbel and Seán O'Leary



Summary

This internship report deals with the synthesis of sound textures based on a statistical description. The synthesis model under investigation was proposed by Josh McDermott and Eero Simoncelli in 2011 (McDermott & Simoncelli, 2011). First the definition of what is a sound texture will be discussed. This is followed by a description of the model under investigation. Insights on the theoretical aspects of the model and on its implementation will be given. The limitations of the algorithm will be discussed and related to the theoretical definition of sound textures. Some proposition to possibly improve the implementation are made. Finally some preliminary work on identifying links between higher-level parameters, such as density, and low-level statistics will be presented.

Keywords : sound textures, statistics, auditory perception

Ce rapport de stage traite de la synthèse des textures sonores basée sur une description statistique du son. Le modèle de synthèse étudié est le modèle introduit par Josh McDermott et Eero Simoncelli en 2011 (McDermott & Simoncelli, 2011). Après avoir essayé de définir plus précisément en quoi consiste exactement une texture sonore, nous détaillerons le fonctionnement du modèle sur le plan théorique. Une description détaillée de l'implémentation sera également présentée. Des propositions pour améliorer l'implémentation seront présentées. Les limites des possibilités de synthèse du modèle seront discutées et mises en relation avec la conception théorique que nous avons des textures sonores. Finalement un travail préliminaire portant sur l'identification de liens entre des paramètres de haut-niveau, tels que la densité d'une texture sonore, et l'encodage statistique de bas-niveau sera présenté.

Contents

1 Introduction

Sound textures synthesis is a rather complex problem, involving different research fields such as signal processing, acoustics and auditory perception. The goal is to synthesize sounds which present internal variations, some textured quality, but which somehow remain similar over time.
Many different approaches to sound textures have been proposed so far. This work focuses on a recent synthesis method, proposed by Josh McDermott and Eero Simoncelli in 2011, which is based on a completely statistical description of sound textures. But the statistics used in the model are abstract and do not provide an intuitive platform for sound synthesis and transformation. An important part of this work is dedicated to understanding what statistics can encode about sound textures, and what they cannot. It seems that some sounds are well synthesized using statistics, some others are not. Does that mean that the sounds that don't fit in a statistical description are not sound textures ?
In order to get more intuitive control over the synthesis model we would need to link the statistics of the model with higher-level physically descriptive parameters, such as density, which would be general and apply to all sound textures. We present some preliminary steps towards identifying those links.

2 What is a sound texture ?

Before going any further into sound texture synthesis it seems a good idea to try to define what exactly we are trying to synthesize.
At first sight sound textures seem to be a rather common and obvious class of sound. One immediately thinks of natural sounds such as rain, wind, water streams, bird songs etc. But we rarely hear only rain, we might more probably hear a storm with rain falling, wind blowing and some thunder cracks. Do all these events put together still create a sound texture ? Walking through the countryside, we could also include less natural sounds like crowds, motor sounds, cars passing on a highway. But if the traffic gets sparser where individual cars can be perceived, can this still be considered a sound texture ? Even if we get some general intuition about sound textures by considering examples, defining more clearly the class of sound textures is necessary if we want to determine wether or not our sound texture synthesis is "good". But by gaining in abstraction and formalization we also narrow our idea of what a sound texture is and reject a lot of potentially interesting candidates. As Strobl and Eckel state, " there is no valid universal description of what is a sound texture" (Strobl & Eckel, 2006). Daniel Mölhmann also suggests that there exist two ways of dening sound textures, the "narrow" definition and the "wide" definition (Mölhmann, 2011, pp. 18-19). "Narrow" deals only with simplistic sound textures which are very dense and noisy, like rain. "Wide" definition includes sounds that have much more complex high-level patterns, e.g. Lu et al. (2004) or Misra, Cook, and Wang (2006)).

Most researchers on sound textures often agree on defining sound textures as a sound that is made of a random superposition of acoustical events that exhibit some sort of similarity. This concept was first introduced by Nicolas Saint-Arnaud (Saint-Arnaud, 1995). Saint-Arnaud suggests that sound textures are presented as built on two levels. A low-level which consists of sound atoms. These are basically the single acoustical events e.g. a rain drop, a car passing, a voice in the crowd. The high-level refers to the organization of these events, i.e. their distribution over time. These distributions will most likely be driven by stochastic processes. This model of sound texture, although compact and elegant, seems in certain cases less applicable. For example, the sound of a water stream: how can individual atoms be defined ? Is it realistic to say that water flowing is an aggregate of water drops ? In the case of wind, the notion of individual atoms is also ill-defined. Of

course it is because wind makes no sound, we hear it as its interaction with other elements, such as leaves. But wind is also commonly identified with the whistling sound that is produced when it passes through narrow openings producing resonances, or by the aerodynamic interactions with the buildings. In such cases, the atomic description seems less pertinent.
The two other very important concepts introduced by Saint-Arnaud are the notions of constant long-term characteristics and attention span. Constant long-term characteristics refers to the fact that properties of the sound texture must remain constant over time, although it is not specified which properties exactly. We can understand it as statistical homogeneity in the distribution of the atoms over time, but also as a similarity between each atom. This idea is well described by the curves of figure [1]. The constant long-term characteristics of sound textures implies that the information conveyed comes to a saturation point after a certain amount of time. This is in contrast to signals such as music and speech, which typically have constantly evolving semantic content (if we do no not consider minimalistic or repetitive music), and has a potential for constantly increasing information.

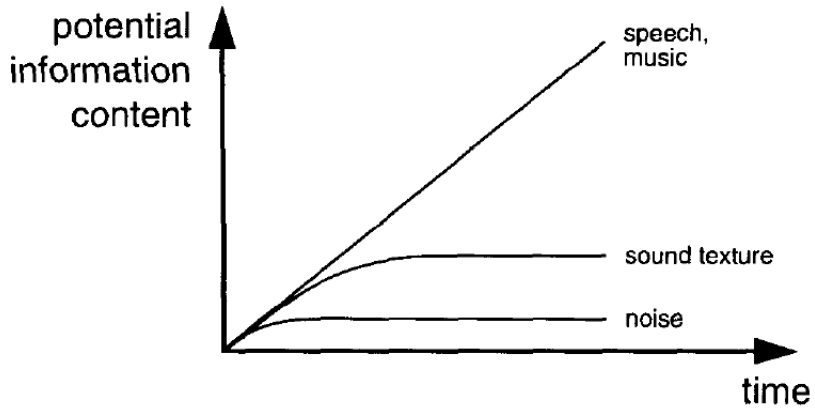


Figure 1: Illustration of constant long-term characteristics for sound textures (image from Saint-Arnaud)

According to Saint-Arnaud, attention span is "the maximum time between events before they become distinct". The previous example of the highway is a good example. If cars get too sparse they will not be perceived as a whole but as distinct events and therefore not as a sound texture. But it does not

mean that we cannot perceive individual events in a sound texture. In his attempt to derive a general typology for sounds, Pierre Schaeffer delivers some interesting thoughts about what he already considered as sound textures : " It only depends on our own will to hear the sounds of gravels falling from a truck as coming from a unique source (the action of the truck), or as made out of multiple sources (each gravel falling on another)."[1] (Schaeffer, 1966). This quote perfectly summarizes all the perceptual ambiguity of sound textures.

For the rest of this report we will try to stick to Saint-Arnaud's definition since it is the most commonly used and it is also quite general. But we shall keep in mind that sound textures are complex objects and that maybe no definition is general enough to encapsulate all of their subtlety.

[1]"En effet il dépend de notre volonté d'entendre le son d'une coulée de cailloux comme provenant d'une cause unique (la benne qui se déverse), ou comme composée d'impulsions brèves dues à une multiplicité de causes analogues (chaque caillou tombant sur les précédents)."(Pierre Schaeffer, 1966, *Traité des objets musicaux ,p. 454*)

3 Overview of the algorithm under investigation

Since the pioneering works of Saint-Arnaud and Popat in the mid 1990's, sound textures synthesis has been capturing more and more of the audio analysis and synthesis community's interest. Many different models have been proposed, using such varied methods as wavelet tree learning (Dubnov et al., 2002), time-frequency LPC (Athineos & Ellis, 2003, Zhu & Wyse, 2004), and granular synthesis (Hoskinson & Pai, 2001, Schwarz et al., 2006). For the purpose of this work, we will focus on a new recent and promising model, based on statistical analysis informed by the response of the peripheral auditory system (McDermott et al. 2009, McDermott & Simoncelli, 2011). This model is based on the idea that our perception of sound textures can be described by a certain well-chosen set of statistics. If this is the case, given a sufficient set of statistics it should be possible to re-synthesize a perceptually equivalent texture by constructing a sound with the desired statistics. This logic is inspired by research on visual textures (Portilla & Simoncelli, 2000).

McDermott's model is particularly interesting for two main reasons. Firstly, because it doesn't use sampled sound but generates sound textures from scratch, by iteratively imposing the right statistics on white noise. The only information we need is the statistics of the desired sound texture. Moreover, since the synthesis is based on statistics, we can generate examples of any length. This is potentially beneficial when applied to domains where we have strong constraints on the storable amount of data, such as video games. The second reason is that it produced very compelling results on varied sound texture examples such as water, waves, wind, insects and even mechanical sounds. We are thus tempted to think that this model constitutes a quite general model for sound texture synthesis.

It is however important to note that McDermott's approach differs from other works on sound texture in that sound synthesis is not the ultimate goal but is used as a mean to demonstrate certain hypotheses about the perception of sound textures. As he explains in his 2011 paper : "Our goal in synthesizing sounds was not to render maximally realistic sounds [...] but rather to test hypotheses about how the brain represents sound texture".

The first stage of the model is based on the processing of the auditory periphery. An ERB filter bank and non-linear compression are used to simulate the

processing of the basilar membrane. The basic hypothesis of the synthesis model is that simple statistics from the early stages of auditory processing are used at a later stage to recognize and distinguish sound textures. In his latest researchMcDermott demonstrates the sufficiency of time-averaged statistics for texture recognition and categorization (McDermott, Schemitsch & Simoncelli, 2013). The underlying idea is that when listening to sound textures we do not focus on individual events but rather on a global behavior of the sound events, which can be coded by time-averaged statistics. One might think about thermodynamics and the use of statistical descriptions for the diffusion of gases, rather than modeling the trajectory of each particle.

We will now enter technical explanations of the statistics used in the model. We will then discuss the implementation of this model, including the method used to impose the desired statistics on the synthesized sound. We believe that it is important to separate the general idea of the model, based on strong perceptual assumptions, from its actual implementation. The synthesis relies mainly on conjugate gradient descent to compute the envelopes. This method does not guarantee that the resulting envelopes will have the desired statistics. Because the synthesized sounds cannot be expected to have the exact statistics desired it can be unclear if the loss in quality of the synthesis is due to a perceptually inadequate description of the signal or because the gradient descent did not find an optimal solution.

3.1 Statistics of the auditory periphery

The philosophy of Josh McDermott's algorithm is to compute a set of statistics from a recorded sound texture and then to create a synthetic sound with matching statistics. We are thus actually performing a type of re-synthesis of the original sound, based upon statistical data.
The choice of statistics were informed by the response of the peripheral auditory system. There are two stages involved in the computation of these statistics. The whole process is summarized in figure [2].

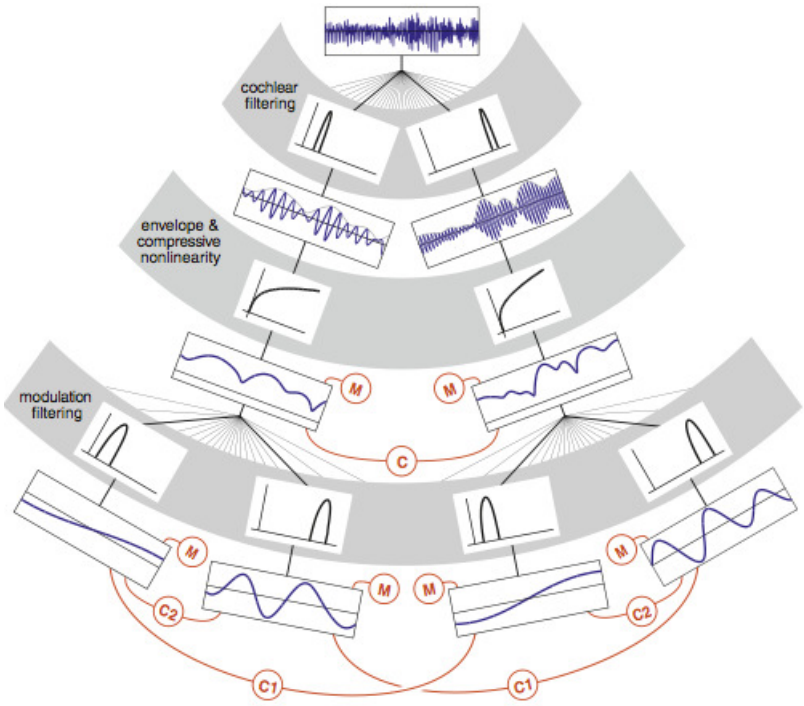


Figure 2: Stages involved in the computation of the statistics (image from McDermott).

In the first stage the sound signal is processed through a bank of 30 band-pass auditory filters that decomposes the sound into acoustic frequency sub bands. These filters are cosine-shaped filters spread on an ERB scale. Each filter overlaps the adjacent one by 50%. The envelope of each sub band is then extracted using a Hilbert transform, these envelopes are then further processed using a nonlinear compression, informed by auditory perception. We shall call the resulting envelopes the sub band envelopes. The model's statistics are calculated on these sub bands. The set of statistics used to describe the temporal trajectories of these envelopes and their interdependence are first-order moments (M) - mean, variance, skewness and kurtosis - and correlations (C).
Noting $s_k(t)$ the envelope of the k-th sub band the moments are defined as follows :

Mean :

$$M1_k = \mu_k = \sum_t w(t) s_k(t)$$

Variance :

$$M2_k = \frac{\sigma_k^2}{\mu_k^2} = \frac{1}{\mu_k^2} \sum_t w(t)(s_k(t) - \mu_k)^2$$

Skewness :

$$M3_k = \frac{1}{\sigma_k^3} \sum_t w(t)(s_k(t) - \mu_k)^3$$

Kurtosis :

$$M4_k = \frac{1}{\sigma_k^4} \sum_t w(t)(s_k(t) - \mu_k)^4$$

A windowing function $w(t)$ can optionally be used. In our experiments we used a rectangular window (i.e. a constant scaling factor, effectively no windowing). The four first-order moments describe the energy distribution of the texture. If we had envelopes with gaussian distribution for example, the sole knowledge of the mean and the variance would be sufficient to re-construct the original texture. But textures are more complicated than gaussian noise, hence the use of skewness and kurtosis for a more complete description.

3.1.1 An Intuitive Discussion of the Statistical Description

In order to get some intuition about what each statistic is coding, we show spectrogram examples of the re-synthesis of a texture made of periodic broadband noise bursts (figure [3], top), each time omitting a different statistic.
Moments capture information about the shape of the envelope histogram, i.e. the distribution of the energy. The variance describes the degree to which the amplitude of the envelope is not close to the mean. Higher-order marginals characterize both the peakedness and the sparsity of the sound texture. For example the presence of a certain events of amplitude much greater than the mean will tend to increase the skewness of the distribution. If we get many of these events then the mean will increase. High kurtosis may reflect the presence of impulsive sound events (Erdreich, 1986), whereas low kurtosis may characterize slowly modulated sounds.

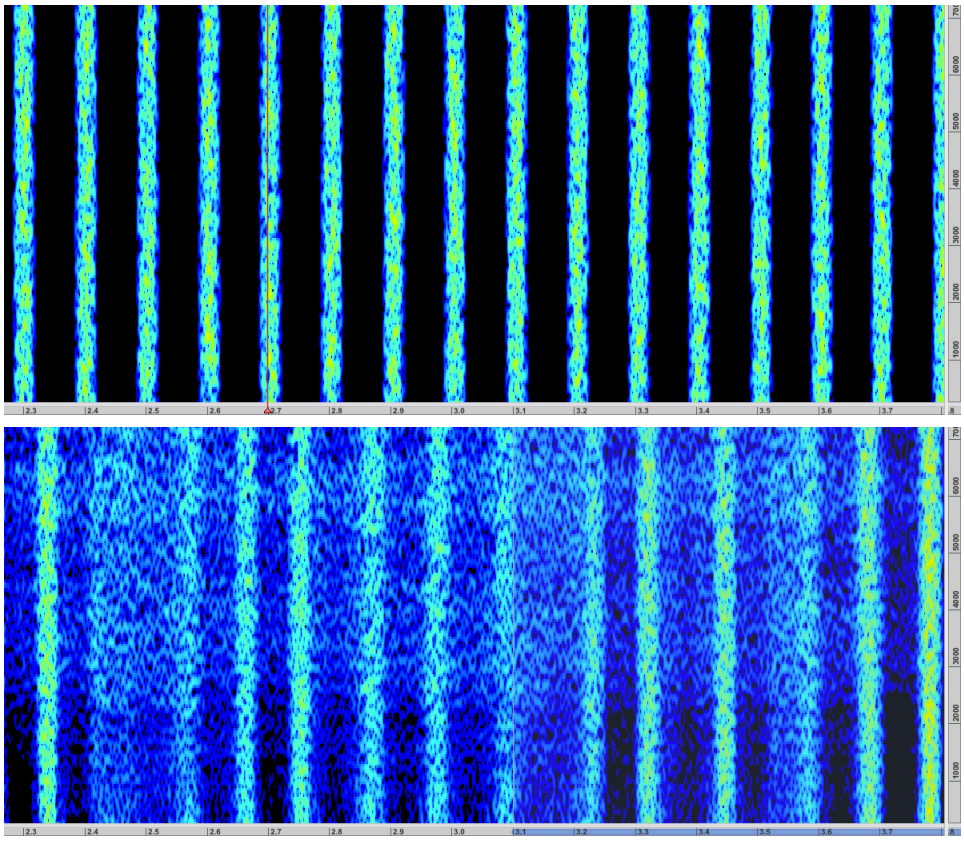


Figure 3: Re-synthesis of a stream of periodic broadband noise bursts, omitting variance, skewness and kurtosis. Top is original, bottom is re-synthesis.

Correlations between envelope sub bands are computed as :

$$C_{jk} = \frac{1}{\sigma_j \sigma_k} \sum_t w(t)(s_j(t) - \mu_j)(s_k(t) - \mu_k)$$

where σ_j is the standard deviation of $s_j(t)$. The correlation of the sub bands encodes the degree of temporal synchronicity of the envelopes. This can be seen as describing the vertical structure in the spectrogram. Correlation between envelopes is not sufficient to capture the frequency structure of the texture. A finer level of correlation is thus introduced : C1 correlations. In the second stage each of the sub band envelopes are processed through a bank of bandpass modulation filters. In the case of C1 correlations, a bank of 6 octave-spaced filters is used. Their center frequencies vary from 3 Hz to 100

Hz. These filters overlap by 75%. The resulting signals are called modulation bands. C1 correlations are correlations between similar modulation bands of different sub band envelopes. Noting $\hat{b}_{j,n}$ the n-th modulation band of the j-th auditory band, C1 is computed as :

$$C1_{jk} = \frac{1}{\sigma_{j,n}\sigma_{k,n}} \sum_t w(t)\hat{b}_{j,n}(t)\hat{b}_{k,n}(t)$$

where $\sigma_{j,n}$ is the standard deviation of $\hat{b}_{j,n(t)}$. There is also another type of correlation, C2 correlation, which consists of correlations between different modulation bands on the same sub band envelope. Their role is to capture the phase relationships inside a certain sub band. Their computation is slightly more complicated since it involves squaring analytic versions of the neighboring modulation band in order to double its frequency and thus compare the phase. Details can be found in the original paper of McDermott and Simoncelli (McDermott & Simoncelli, 2011). In practice, the use of this statistic didn't seem to improve the synthesis and we omitted it in most of the experiments.
Figure [4] shows re-synthesis of periodic broadband noise bursts with omitting all the correlations. We clearly see the loss structure with frequency.

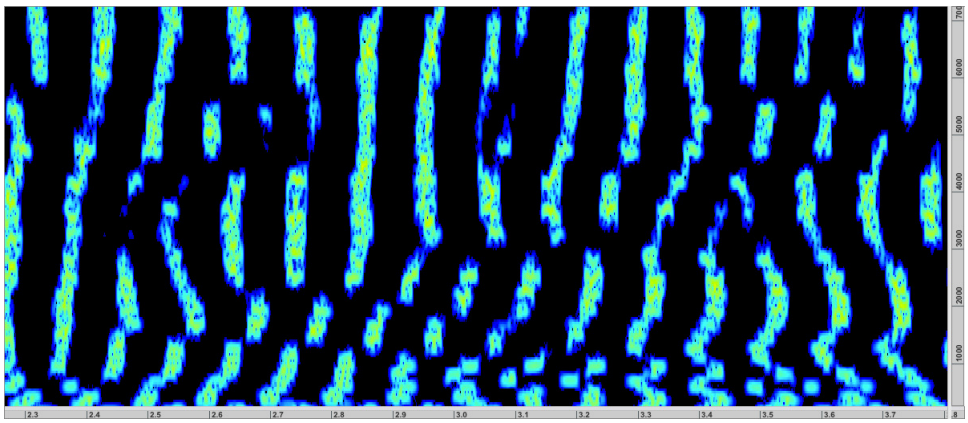


Figure 4: Re-synthesis of a stream of periodic broadband noise bursts, omitting all correlations.

The last type of statistic used is the modulation power. Modulation power is the variance of the modulation bands. The modulation power is computed using a different filter bank to that used for C1 and C2 correlations. The

bank consists of 20 bandpass modulation filters. These modulation filters are tuned to capture slow rate modulations; their center frequencies vary from 0.5 Hz to 200 Hz with logarithmic spread. They also overlap by 75%.
Noting $b_{k,n}$ the n-th modulation band of the k-th auditory band, modulation power is computed as :

$$MP_{kn} = \frac{1}{\sigma_k^2} \sum_t w(t) b_{k,n}(t)^2$$

where σ_k is the standard deviation of $s_k(t)$, the corresponding sub band envelope.
Modulation power being related to the energy of each modulation band, it can be viewed as a logarithmically sampled power spectrum (with some smoothing due to the overlapping filters) for each sub band envelope. The Wiener-Khinchin theorem stating that power spectrum is the Fourier transform of the autocorrelation function, modulation power is related to the autocorrelation function and describes the time structure of each sub band envelope.

3.2 Current implementation

3.2.1 Gradient descent

In order to obtain envelopes with the desired statistics gradient descent is used. Gradient descent is a technique used for optimization problems. Given a certain constraint function f and an initial vector x_n, we want to find a new vector, x'_n that minimizes f. We can proceed by iteratively imposing:

$$x'_n = x_n - \gamma \frac{\partial f(x_n)}{\partial x_n}$$

x'_n being the new value of the vector after the next step of the gradient descent and γ a parameter related to the step size of the descent. Geometrically, the algorithm is considering the constraint function as an N-dimensional surface. Starting from a point defined by the initial 'guess' x_n (in our case one of the sub band envelopes of the initial white noise), the algorithm tries to move towards a local minimum in the function. The direction chosen by simple gradient descent is that of the direction of the negative gradient. This is lo-

cally the direction of steepest descent. The choice of the step size determines the rate of descent and can strongly effect the solution obtained. Too small a step size and the algorithm will converge to a solution slowly, too large and a true minimum may never be reached. With an appropriate step size eventually we are sure to come close to a minimum, but we have no guarantee that it will be a global minimum.
Carl Rasmussen's "minimize" MATLAB function is used for all gradient descent calculations.

In the current implementation of the algorithm, the minimized function $D\left(S, \hat{S}\right)$ is the sum of the squared differences between the desired set of statistics S_i and the current set $\hat{S}_i$. Index i denotes an individual statistic from the set (moment, correlation,...).

$$D(S, \hat{S}) = \sum_{i \in stats} (S_i - \hat{S}_i)^2$$

The desired set of statistics S_i is computed from the original texture and is constant. The current set $\hat{S}_i$ is computed for each sub band envelope. The sub bands envelopes are not updated altogether but one after another, in order of decreasing energy, and an independent gradient descent is performed for each of them.
Naming x_n a certain sub band envelope, the gradient of $D\left(S, \hat{S}\right)$ regarding this sub band envelope is :

$$\frac{\partial D(S, \hat{S})}{\partial x_n} = \sum_{i \in stats} -2(S_i - \hat{S}_i) \frac{\partial \hat{S}_i}{\partial x_n}$$

Therefore the gradient at each iteration is entirely determined by the derivative of each of the statistics regarding the sub band envelope being optimized. Also the gradient is not dependent on the sub band being considered. After each iteration x_n is slightly modified so that $\hat{S}_i$ gets closer to S_i. $D(S, \hat{S})$ always being positive (sum of squared real numbers), the convergence to the global minimum would imply that we are exactly matching the desired statistics for this sub band, assuming such a solution exists.

3.2.2 Imposition of the statistics

In this section we will discuss how gradient descent is used to impose the desired statistics on the sub band envelopes. Here iteration will refer to an iteration of the global process of the algorithm and not to a step of the gradient descent.
In order to impose the desired statistics on each sub band envelope the envelope of each sub band signal is separated from the phase (also called "fine structure"). The envelopes are extracted by taking the absolute value of the analytic signal (the analytic signal is generated using a Hilbert transform). The phases are obtained by dividing the analytic signal by the envelope. Gradient descent is then used to impose the desired statistics, as described above. However, only a few steps are taken in the gradient descent (between 5 and 15) at each iteration. Thus after an iteration of the algorithm the statistics are only partly imposed. This is because the re-combined signals (envelopes and phases) are no longer guaranteed to be band limited after the gradient descent. And so after each iteration the updated envelopes are re-combined with the phases (non-updated) to create updated sub band signals. These are then re-filtered using the original auditory filter bank (these filters have a half cosine frequency response, two passes gives a cosine squared response. With 50% overlap the combined frequency response of this filter bank is flat). The implementation is represented in figure [5].

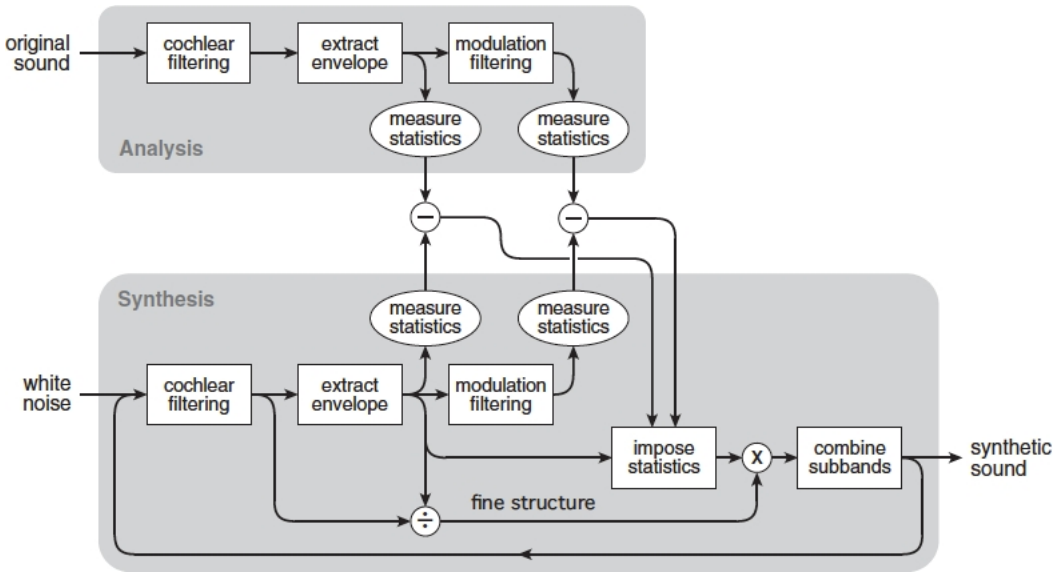


Figure 5: Current implementation (image from McDermott).

This is a very awkward choice of implementation because by filtering we lose a little of the optimization made during the gradient descent. Also it is very strange to re-introduce the phases since the model is entirely based on statistics of the envelopes. And so this implementation compromises the optimization process.

Introducing another way of imposing the statistics is not such an easy thing to do. The main problem is that the bandwidth of the sub band signals cannot be guaranteed to remain band limited after gradient descent. We can't impose bandwidth constraints on the envelopes because they are computed by taking the absolute value of an analytic signal, they can have discontinuities. However we intuitively understand that, apart from a supposedly small number of points of discontinuity, the envelope of a low frequency sub band with narrow bandwidth is necessarily smoother than that of a high frequency sub band with large bandwidth.
One way to implement the imposition of the statistics in one single gradient descent, without intermediate merging and splitting, would be to have the envelopes down sampled at different rates. The Nyquist frequency would thus impose the sub band envelope frequency limit. But we would then have

to deal with multi-rate signals, which would greatly complexify the computation of statistics, particularly the correlations.

Another way that would avoid multi-rate computations would be to use a signal model for the envelopes, and compute the gradient not regarding the samples of the envelopes, but regarding the parameters of the model. A b-spline model would here come handy because it yields easy relations when deriving regarding the spline parameters. If we define envelopes x_n as splines of basis $B_m(n)$ and of M coefficients α_m we have :

$$x_n = \sum_{m=1}^{M} \alpha_m B_m(n)$$

For $m \geq 3$ (i.e. using at least quadratic b-splines), these envelopes will be band limited by the number M of breakpoints.
Then the gradient of the statistics of the sub band being updated is :

$$\frac{\partial \hat{S}_i}{\partial \alpha_m} = \frac{\partial \hat{S}_i}{\partial x_n} \frac{\partial x_n}{\partial \alpha_m} = \frac{\partial \hat{S}_i}{\partial x_n} B_m(n)$$

So it is just the gradient of the statistics regarding the envelope (that we already know) multiplied by the basis function.
Unfortunately due to time limitations those methods haven't been implemented or tested. It should be noted that while using such a signal model for the envelopes may band limit the envelopes during gradient descent, it does not guarantee that the sub band signals will be band limited when the envelope is re-combined with the phase. Also even if imposing the statistics in one single gradient descent is neater, maybe slightly faster, and allows more control over the algorithm, it may not yield better convergence of the statistics, or better sounding results.

4 Synthesis limitations

McDermott's algorithm provides to this date the most compelling sound texture synthesis without using original sound samples (such as in granular synthesis). However there are still many cases where the algorithm yields results that are not very convincing, or really far from the original examples. This algorithm seems to be particularly efficient in the case of very noisy and dense textures, thus implicitly fitting to the "narrow" definition of sound textures, but comes more unstable in other cases.
Sound examples would be more representative but as a substitute we use spectrograms to compare original sounds and re-synthesis.

4.1 Limitations due to frequency resolution

Due to the logarithmic spacing of the auditory filters, there is a larger bandwidth at high frequency. Fine frequency structure, such as a resonance cannot be captured with those filters. This is demonstrated on the example below (figure [6]). The sound texture is made of small cereal grains being poured in a cup. The resonance of the cup can clearly be seen on the spectrogram of the original sound. After re-synthesis, the frequency structure is completely lost. Only remains a resonance around 300 Hz. Indeed filters being narrower at low frequencies, low frequency resonances are still captured.
For the same reasons of frequency resolution of the model, phenomenons such as chirps cannot be captured. This is particularly obvious in the example of a sound texture made of insects sounds (figure [7]).

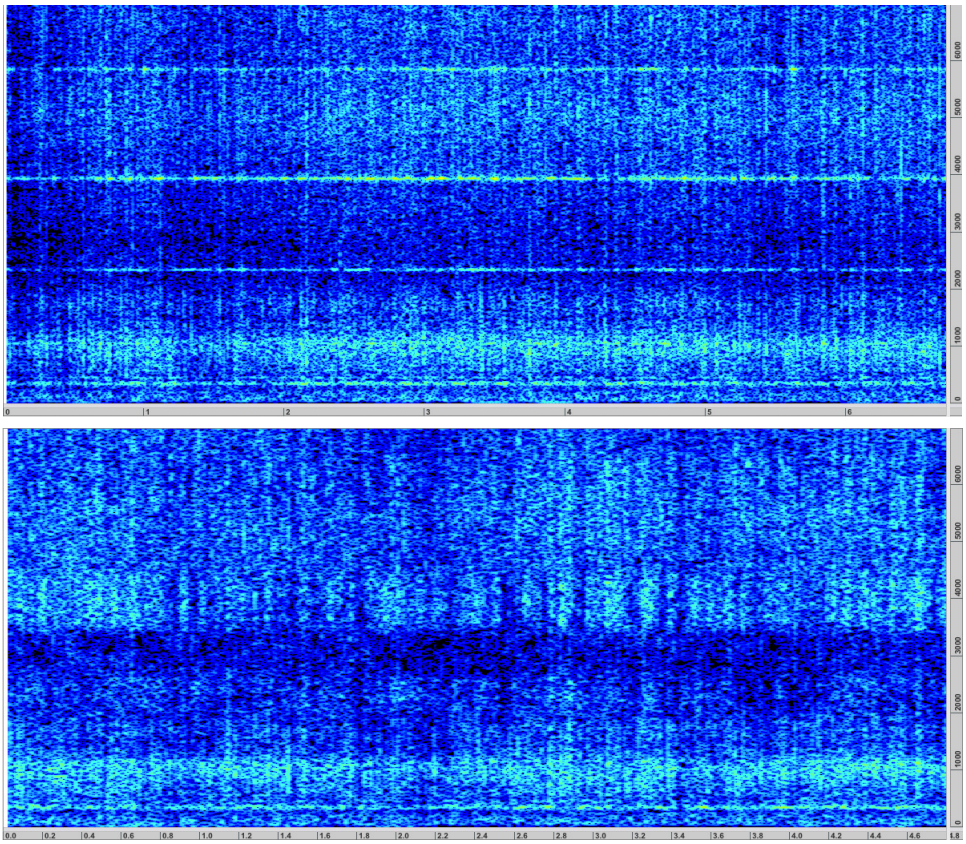


Figure 6: Example of re-synthesis of a sound of grains falling in a cup. Top is original, bottom is re-synthesis.

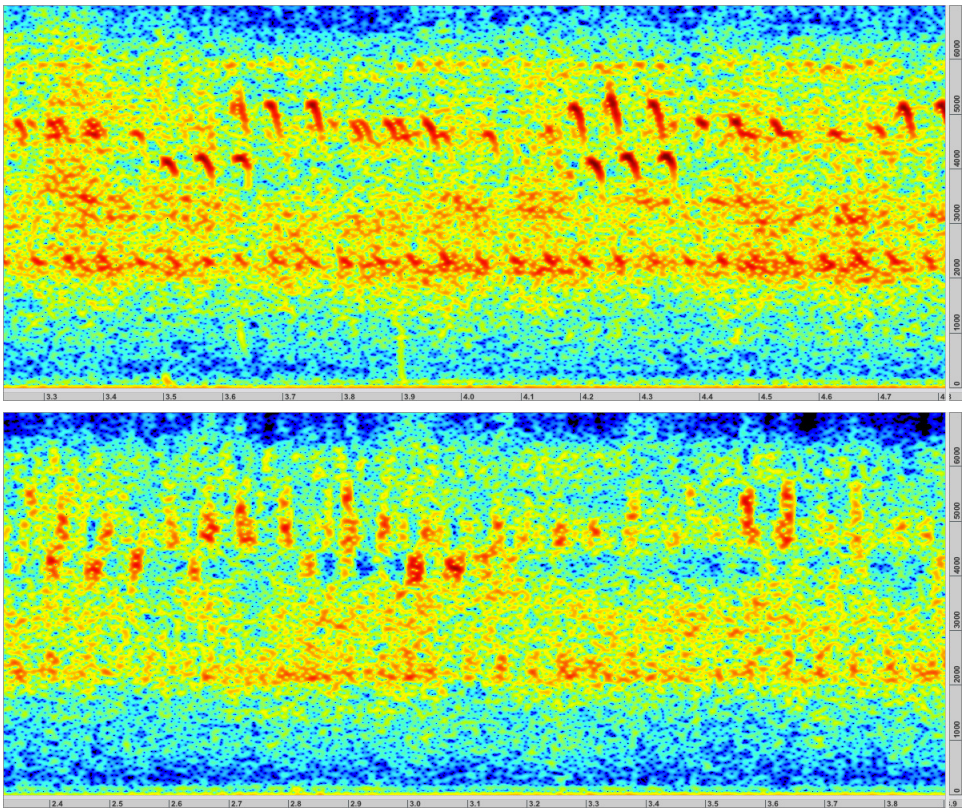


Figure 7: Example of re-synthesis of an insects sound texture. Top is original, bottom is re-synthesis.

We clearly see on the previous example that the chirped shape of individual events is lost after re-synthesis. The re-synthesized events are just bursts of noise.
What is surprising about this example is that even if we can clearly see the degradation introduced by the algorithm, the re-synthesized texture sounds very close to the original. This is the real power of the algorithm : even if the synthesized sounds are only made of amplitude modulated noise, the use of perceptual assumptions ensures that our perception will get "fooled". However when listening carefully we can clearly hear the difference. We are not able to distinguish individual events from the mass in re-synthesized texture (we are for example unable to count the number of chirps, which we could do with the original texture).

4.2 About homogeneity

It seems that the sounds the algorithm failed most at re-synthesizing were sounds that were statistically varying over time, i.e. sounds that were not homogeneous. The sound texture considered here consists of gravels falling. The re-synthesis yielded a very bad sounding result. The loss of time and frequency structure can be seen on the spectrograms of figure [8].

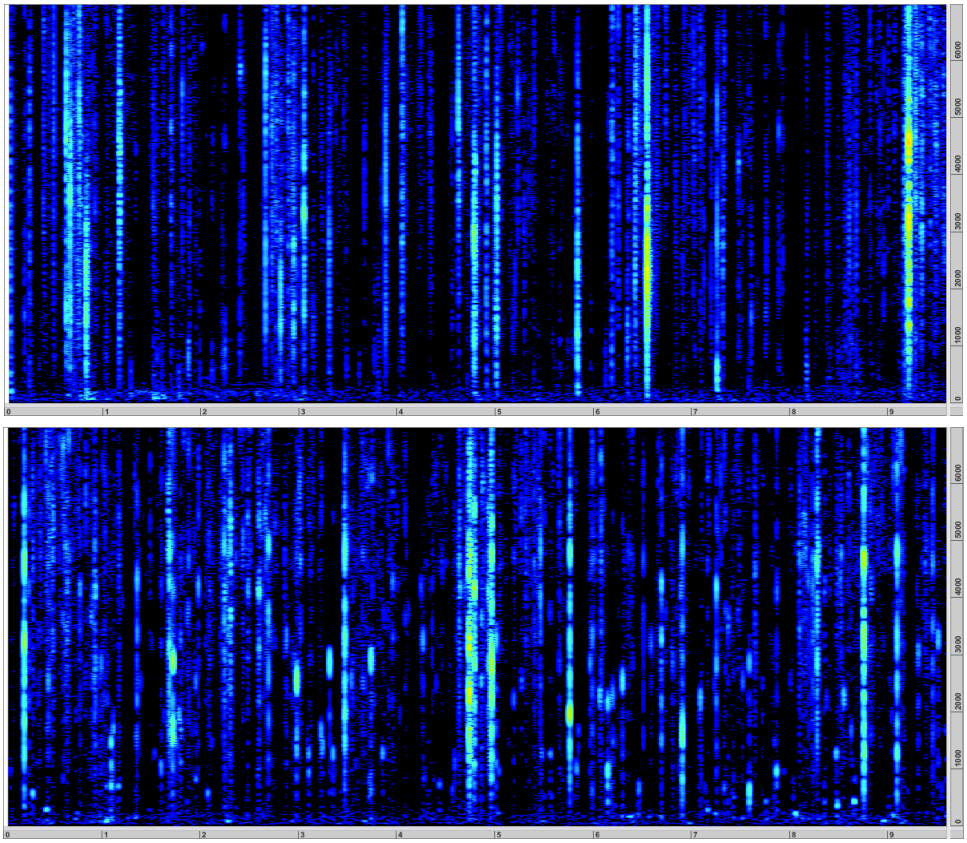


Figure 8: Example of re-synthesis of a gravels sound. Top is original, bottom is re-synthesis.

In order to demonstrate that the statistics are varying over time we split the original sound into smaller segments of constant length and compute statistics on each of these segments. On figure [9] we plot moments over different time segments for the gravel sound. Each blue curve is a moment

over a certain segment. If we compare with figure [10], where is plotted the same analysis for a fountain sound, we can see that we have much more variability for the gravel sound. Such behavior is also found when looking at other statistics. This variability is the sign of a sound that is not statistically homogeneous.

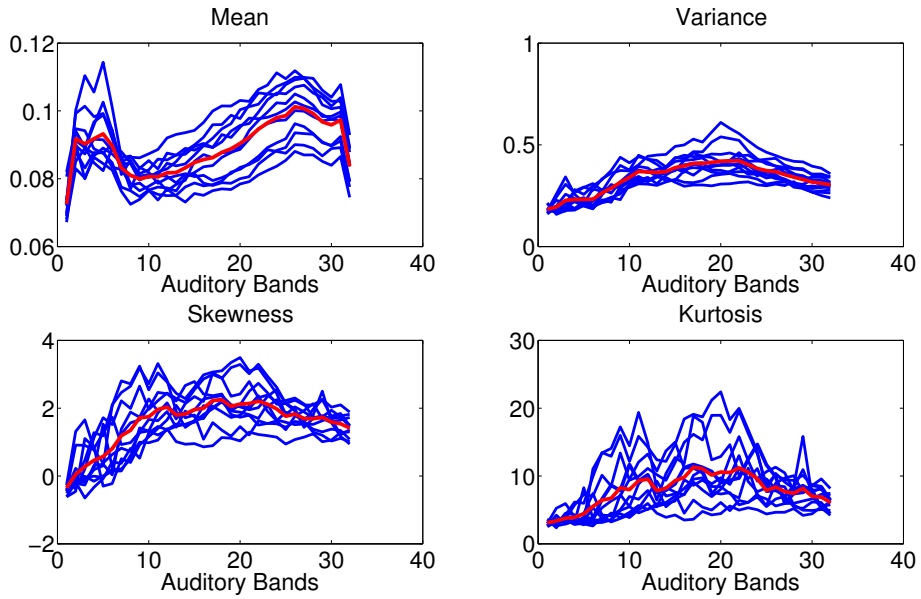


Figure 9: Moments over different time segments, for a gravel sound. There are 11 segments of 1 second, overlapping by 20%. Each blue curve is a moment over a certain segment. The red curve is the mean over all segments.

When measuring the statistics as time-averaged for the whole sound we measure something that is close to the red curve. For a sound that exhibits such variability as the gravel sound, the re-synthesis will necessarily yield a sound that is not close to the original.

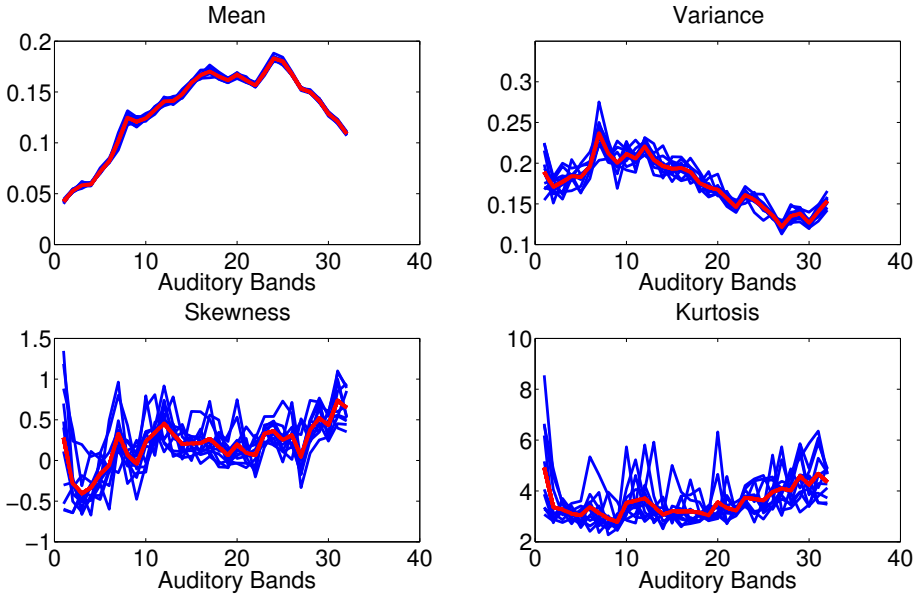


Figure 10: Moments over different time segments, for a fountain sound. There are 11 segments of 1 second, overlapping by 20%. Each blue curve is a moment over a certain segment. The red curve is the mean over all segments.

To get a more precise measure of statistical homogeneity we compute the variance of a statistic (here we choose modulation power) over time segments as a function of the length of the segments. We do it for three sounds: gravels, fountain and pink noise. The idea is to compare at which rate the variance increases for sounds of different homogeneity. Indeed there is a natural increase of the variance when the number of samples (e.g. the segment length) gets smaller. This is why pink noise is chosen as a reference for statistical homogeneity.

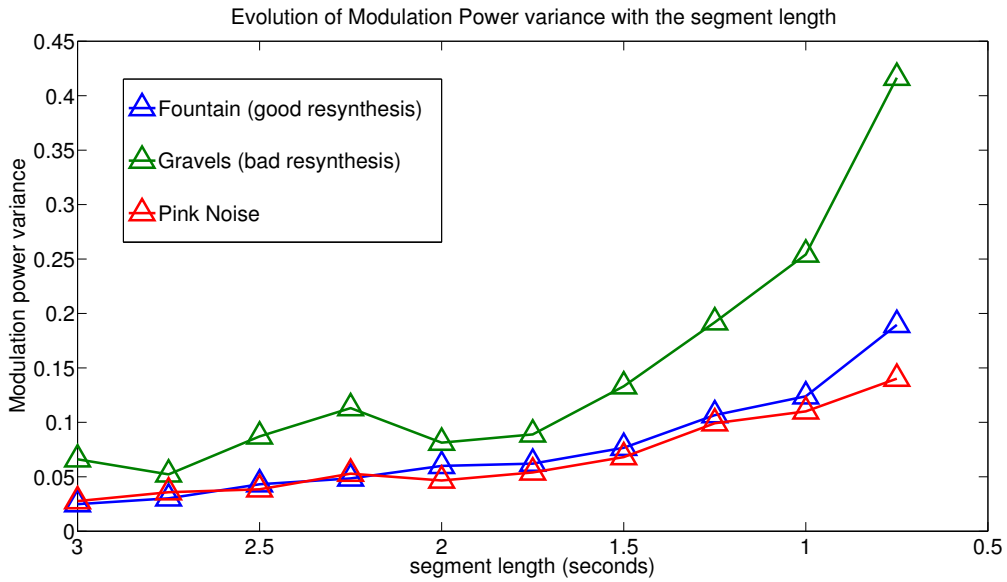


Figure 11: Variance of the modulation power over time segments as a function of the length of the segments.

The results are quite clear. The fountain sound evolves similarly to pink noise for all but the shortest time scales. The gravel sound appears as more heterogeneous at each time scale.
Homogeneity is often considered as a defining feature of sound textures. This is what Saint-Arnaud describes as constant long-term characteristics (cf first section). We are however convinced that the sounds we are dealing with here clearly belong to the category of sound textures, if such a category does exist. The gravel sound is a sound texture even if it is not homogeneous. This might be evidence that time-scale could have an important role in the definition of what a sound texture is. As seen on figure [11] the gravel sound seems to be homogeneous for long segments (large time-scale). But its variance increases exponentially at short time-scale. This would suggest that sound textures are a multi-scale phenomena. Maybe including statistics of statistics (such as the variance of certain statistics) to the set of statistics used in the algorithm would allow to re-synthesize textures with a more complex high-level structure such as the gravel sound.

5 Understanding statistics : building the intuition

As we've seen previously the present algorithm provides mainly a framework to re-synthesize an existing sound texture. Our ultimate goal would be to find meaningful ways of manipulating the statistics to induce modifications of a sound texture. Our work constitutes only a preliminary step towards this direction. In the next section a preliminary look at the possibility of modifying sound textures via their statistics is presented.

5.1 Interpolation between sound textures.

5.1.1 Motivations

Finding relationships in a set of 1515 statistics is not something we can figure out at first sight. As a first approach, interpolation between two different textures seems to be an attractive way to modify a sound texture without a thorough understanding of the statistics. In fact interpolation between statistics has already been used in visual texture processing to produce interesting mixes between textures (Portilla & Simoncelli, 2000).

Given a certain set of statistics A_i corresponding to a texture A and B_i another set of statistics corresponding to a texture B, we define a texture C whose statistics C_i are obtained by:

$$C_i = (1 - x)A_i + xB_i, \ \forall i \in stats, \ x \in [0, 1]$$

which is just linear interpolation between A_i and B_i.
The underlying idea of interpolation is that the generated sound texture C would be perceptually "in-between" A and B. However we have to note that there is absolutely no guarantee that similarities in statistics correspond to similarities in perception. Interpolation for modulation power is especially meaningless. If we consider the case of two periodic sound textures of different rate, interpolation will not yield a modulation power with energy located on a modulation band of an intermediate rate, but a modulation power with energy spread on the two modulation bands corresponding to the rates of the original sound textures. The interpolated texture will thus not have an intermediate rate, but an undetermined superposition of two rates.

Before trying to use a more refined interpolation method, we first investigate the simplest case of rain textures, where we interpolate between rains of different densities but having modulation power close to that of noise.

5.1.2 The example of rain

The hard rain sound in this example is statistically very close to complete noise. As an experiment we try to interpolate between a light rain sound where we can hear some diffuse background rain with seemingly sparse drops in the foreground and its spectrally matched noise version. The spectrally matched noise is simply gaussian white noise filtered to match the spectral shape of the hard rain. By using different values of the x parameter we expect to hear the presence of more or less foreground, as we tend towards the lighter rain sound.
Unfortunately the interpolated texture doesn't vary much with the value of x. We are not convinced that what we obtain is an intermediate version between the two sounds, or a rain that varies in intensity, but rather some blurred version of the light rain sound. Rigorous perceptual tests would be necessary to evaluate the pertinence of interpolation in this case and in general.
Moreover even if "good" interpolations were achieved, the main problem of interpolation is that it doesn't help in understanding the relationships between the statistics, or how they might vary with control parameters (physical or perceptual). In the following we shall consider textures with varying density; density typically being a fundamental parameter of textures (Saint-Arnaud, 1995).

5.2 Identifying higher-level parameters.

A way to understand and control sound texture synthesis via a statistical model is too find higher-level relationships between the variation of the statistics and the control parameter. If we could find some general pattern in the way the statistics evolve when a sound is, for example, increasing in density, it may be possible to induce variations of density starting from a given example of sound texture.

5.2.1 Use of principal component analysis

Principal components analysis (PCA) is a well-known method for identifying statistical relationships from experimental data. PCA aims at reducing the dimensionality of the problem by projecting our data on a subspace while keeping the variance of the data distribution. The algorithm successively searches for the dimension along which the data variance is maximum, the objective being to attribute most of the variance to a lower dimensional space.

We started by generating a set of artificially generated sound textures. These textures where made of 5 seconds streams of short filtered noise events with center frequency and bandwidth randomized. A set of 10 sound textures were synthesized, their density, defined as the number of events per second, was increasing from 5 to 160 (details on the artificial sound textures in Appendix 1). We first applied PCA on the set of 10 sound textures, analyzing only the envelope moments over the 20th auditory sub band. We thus had 10 subjects and 4 variables. The results of the analysis are shown on figure [12].

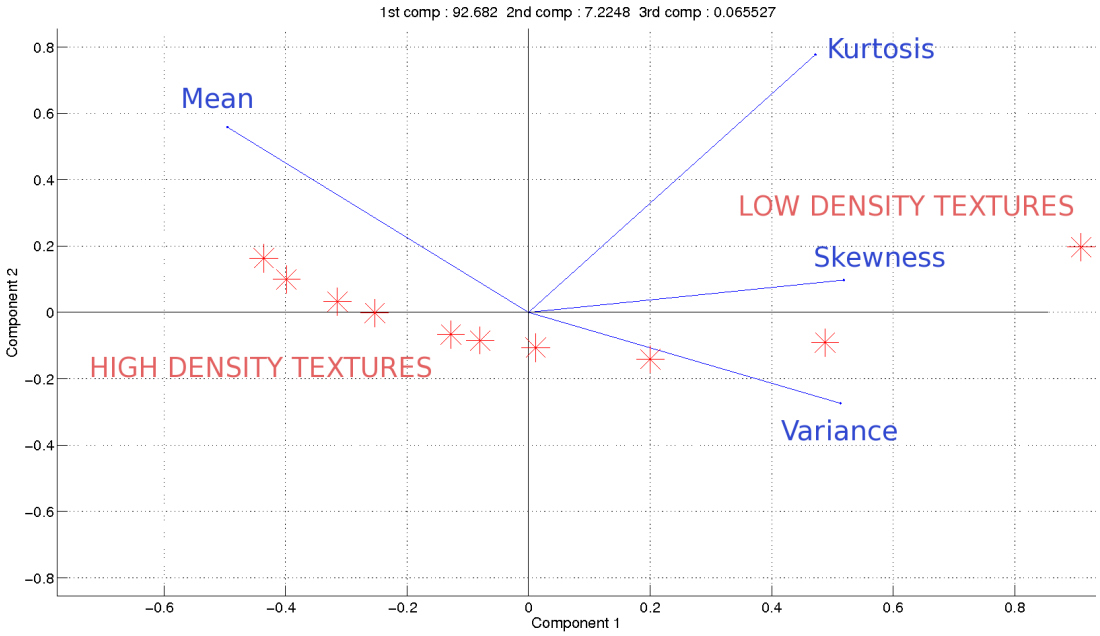


Figure 12: Results of PCA analysis on a set of 10 sound textures (red crosses) of increasing density. The 4 variables (blue segments) are the moments over one band. 99% of the variance of the data is expressed on the first two components.

Once projected on the first two components, subjects are ordered by increasing density, almost following a curve. This analysis confirms the intuition we had in (3.1.1), with some refinement : when textures get sparser their skewness and variance increase. Higher density textures tend to be driven mainly by mean.
We then complete this analysis by including the moments over all the auditory sub bands (figure [13]).

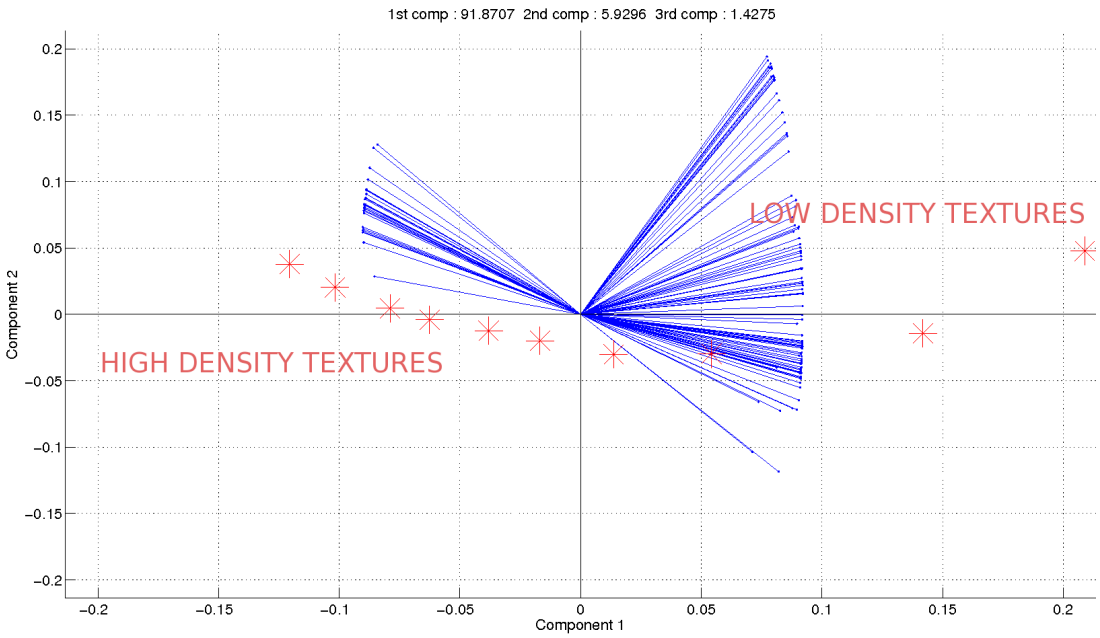


Figure 13: Results of PCA analysis on the same set of sound textures, including the moments for each sub bands (128 variables). Still 96% of the variance of the data is expressed on the first two components.

This analysis confirms the tendency observed in the former, though it is now harder to interpret, since we have a larger number of variables (128). Nevertheless, what we observe is mainly a spread of the variables around the 4 variables of the former analysis. Means over each band and kurtosis over each band tend to be grouped together. Only variance and skewness slightly overlap.
Such analysis thus suggest that we might be able to derive a model for the variation of density of sound textures. Given that principal components are formed of linear combinations of the original variables, if we could fit a curve to the distribution of subjects, we could then directly map variations on this curve to variations of the original variables. However we would have to make sure that PCA explains a sufficient percentage of the original variability of the subjects distribution.

5.2.2 Applying PCA on higher density sound textures

So far we have considered textures of relatively low density (up to 160 events per second). In order to get closer to real life sound textures we now generate artificial textures with higher density (cf Appendix 1). Randomization of the amplitude was also introduced. The objective remains the same : trying to point out relationships between the statistics when a higher level parameter, here density, varies. First analysis only includes moments (figure [14]). The density increases from 5 to 100.

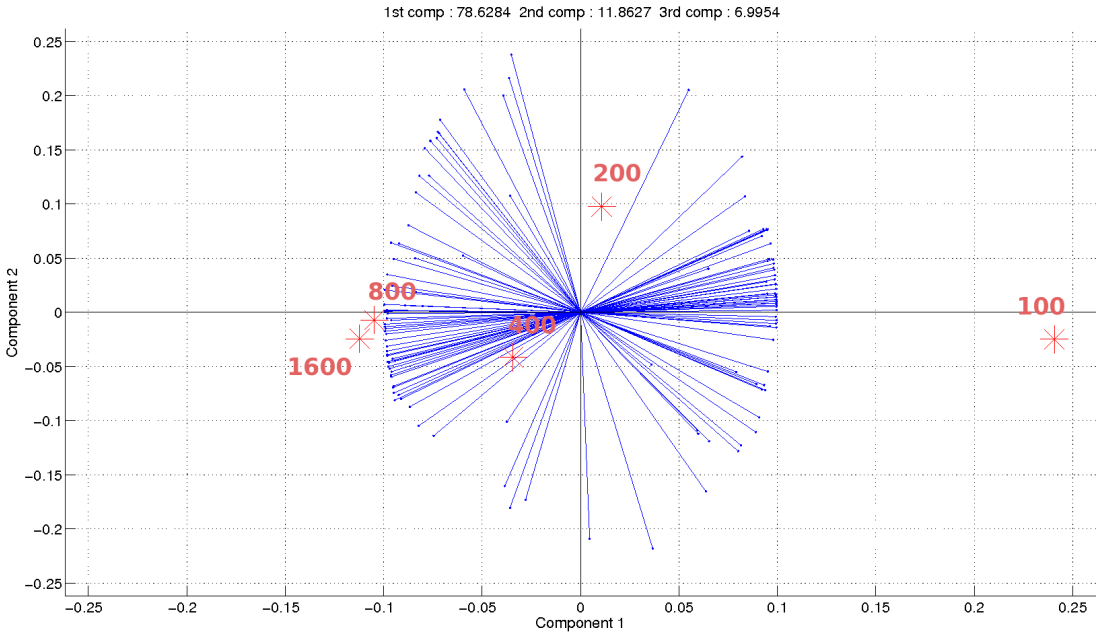


Figure 14: Results of PCA analysis on a set of 5 textures, using the 4 first-order moments. About 90% of the data variance is explained on the first two components. Textures have been annotated with their density.

The results are not as good as for lower density textures. Indeed when dealing with high density textures we have a massive overlap between events. The analysis we made about how skewness and variance coded sparsity is thus less applicable in this case.
Interestingly, if we include more of the statistics in the PCA analysis we can see a tendency in the resulting analysis. The analysis illustrated in figure [15] includes moments (without kurtosis), sub bands envelopes correlations (C)

and modulation power (MP). Kurtosis was removed because it didn't prove to be very relevant for synthesis and it has an almost unpredictable behavior when dealing with high density textures which could upset the analysis. We still have a significant amount of the variance of the data explained on the first two components (83%), which means that we have achieved a drastic dimensionality reduction losing surprisingly little information. The texture distribution could almost be fitted with a linear curve, but the first texture is an outlier.

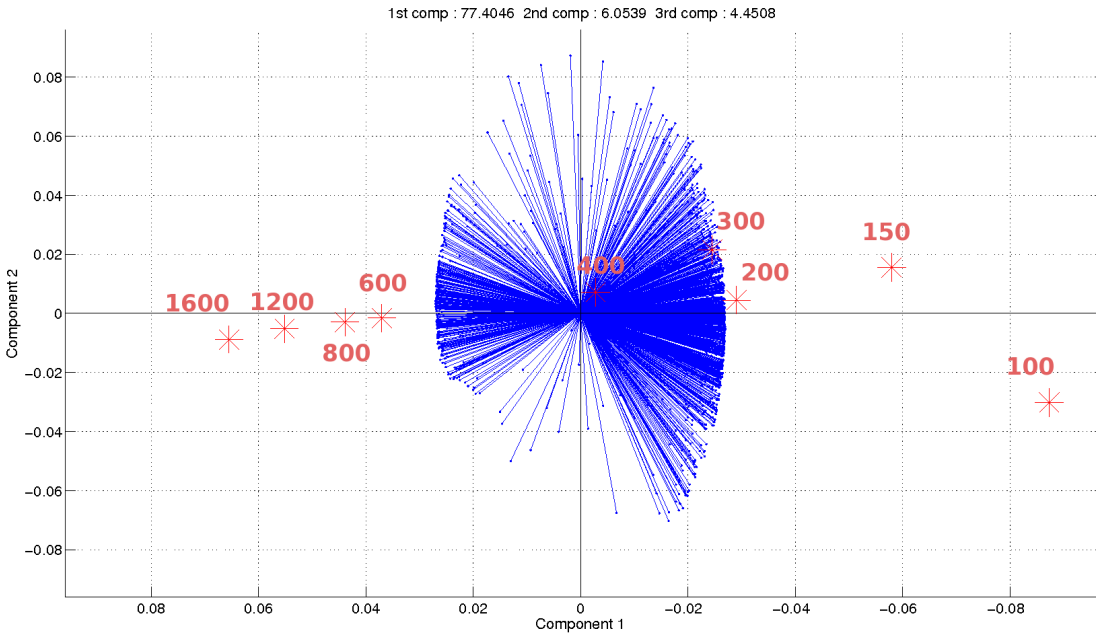


Figure 15: PCA analysis on a larger set of textures, using the 3 first-order moments, envelopes correlations and modulation power for each sub bands . Only 83% of the data variance is explained on the first two components. Textures have been annotated with their density.

The results of this analysis show that the variation of the statistics of the textures under investigation does have some tendency. The moments of the simple textures analyzed (figure [12] and [13]) vary in a predictable way, this show promise for the possibility of modeling of the variation of moments with density. In the following we will investigate the other statistics of the model and their variation with density.

5.2.3 Correlations

We will briefly introduce the variations of correlations with density through a simple example (quasi-periodic texture made of broadband noise bursts). For this example we use a linear filter bank instead of the usual logarithmic filter bank to simplify visualization.

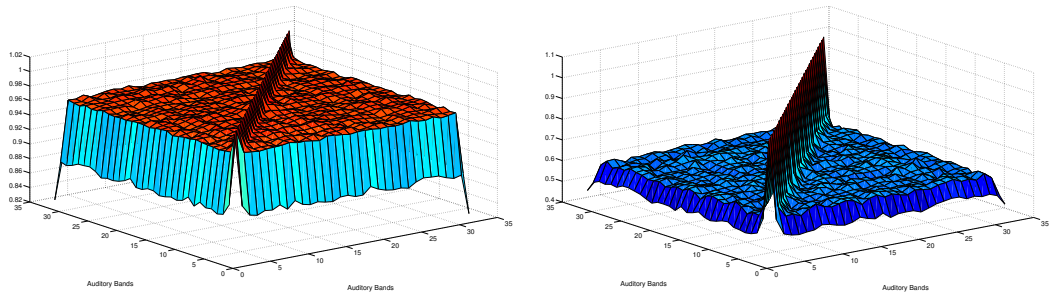


Figure 16: Comparison of correlations for a texture with a density of 4 (left) and a density of 30 (right).

In this example all bands are highly correlated since we are dealing with broadband events, and it makes a flat correlation. Using logarithmic filters, we would have more energy in the high frequency bands. Increasing the density has the effect of scaling down the correlations. Although it may seem counter-intuitive at first sight, because both textures have the same frequency structure, it actually makes sense if we consider that increasing the density is like tending towards white noise. High density textures would then exhibit correlations that are close to that of noise, whereas low density textures would be more correlated.

5.2.4 Modulation power

As discussed in (3.1.1) the modulation power encodes the temporal structure of the sub band envelopes. The evolution of modulation power for an artificial texture (similar to those of used in the PCA analysis above) with density increasing from 100 to 800 events per second is presented in figure [17].

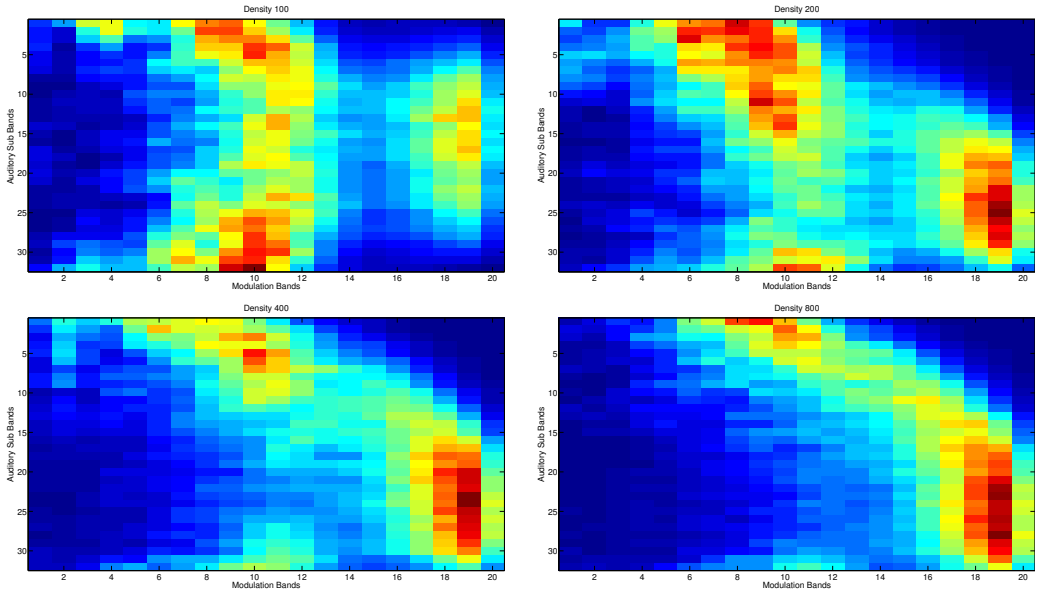


Figure 17: Modulation power for artificial textures with randomized events. These for examples have densities of 100, 200, 400 ad 800 (cf Appendix for details).

Although there is no precise pattern in the evolution of modulation power we can note that at low density the energy seems to rather spread over the low rate modulation bands (left of the picture). When density increases, modulation power progressively shifts to the right, to the high rate modulation bands. In fact, as the density increases the modulation power tends towards that of noise.
The role of the modulation power can be illustrated more clearly with a more deterministic signal. Here we consider the modulation power of signals composed of quasi-periodic broadband events (figure [18]), the rate going from 10 Hz to 30Hz.

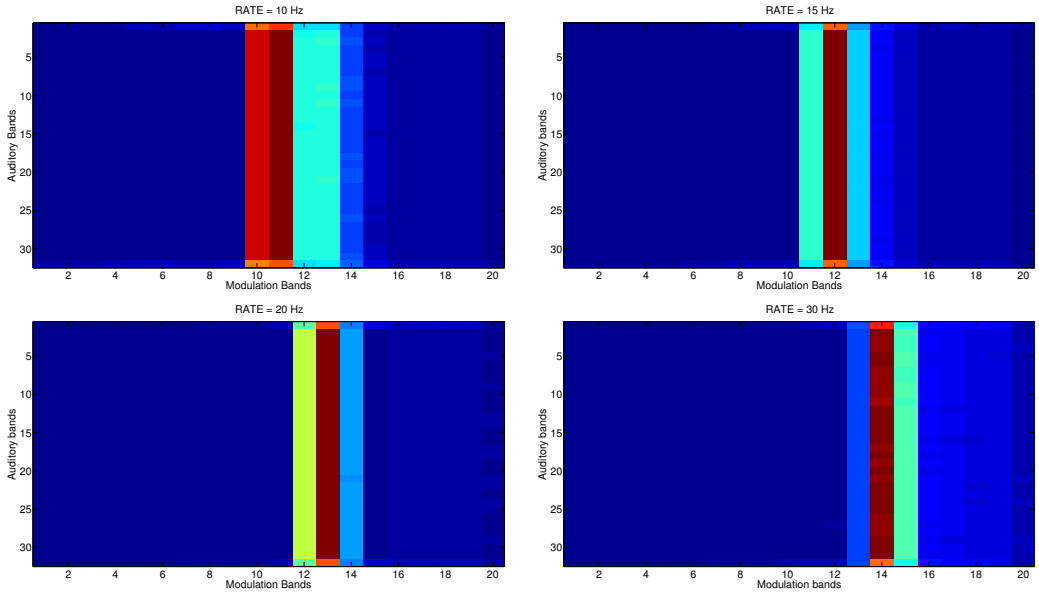


Figure 18: Modulation power for periodic artificial textures with broadband events. Rate increases from 10 Hz to 30 Hz.

We can see that there is a large spread of energy over modulation bands for the low rate texture. The signal being made of quasi-periodic events, its envelope has an almost harmonic spectrum. Since the modulation filters are narrower at low frequency, the presence of higher harmonics in the envelopes induces a spread of energy over the modulation bands. But since the filter gets wider at high frequencies, all the harmonics of the envelopes are covered by a single modulation filter (though there is some spread over the neighboring modulation bands due to the overlap of the filters). Thus rate changes between high-rate textures consists mainly of a shift of the modulation power.

We tried to apply this observation to a quasi periodic texture, in the following case a sample of a helicopter. In order to have control over the rate of the helicopter sound an artificial helicopter texture was synthesized (Appendix 1). For two sounds of close rates, all the statistics but the modulation power remained quite similar. By manually shifting by one band the part of the modulation power which coded the low frequency periodic noise bursts of the helicopter sound and keeping all the other statistics, we managed to synthesize an helicopter sound of lower rate and higher rate. This rate change is

analogous to the pitch shifting of harmonic sounds.
This experiment was performed for small rate changes. For larger rate change we would need to model the change of the spectral envelopes of the modulation power, very much like when pitch-shifting a sound with a phase vocoder. For larger shifts of rate the other statistics will also change significantly. An important conclusion from the above observations is that for textures with randomly distributed events the modulation power tends towards that of noise with increasing density, while for quasi periodic textures the modulation power varies in a way analogous to the pitch shifting of quasi harmonic sounds.

6 Conclusion

This work presented investigations on the use of a statistical description to synthesize sound textures. We have seen that the algorithm we have been using, despite producing very compelling results in some cases, does not cover all the generality of what a sound texture can be.
Limitations in the algorithm may arise from its current implementation, which allows very little control on how the statistics converge. Suggestions have been made on how to improve the implementation, for example using a spline model for the envelopes, though it might be a deeper problem of the combination of phase and envelope. But most probably limitations are due to the use of an incomplete set of statistics. We have seen that sound textures with multi-scale patterns and long-term modulations cannot be correctly encoded with the current set of statistics. Using statistics of statistics might be a possible solution to improve this point while still only using time-averaged statistics.
Synthesizing sound textures from statistics is an interesting approach because it allows one to generate arbitrarily long textures from a restrained number of parameters. This could be very useful for coding purposes. However introducing general meaningful modifications to sound textures from a statistical description is not something we are close to being able to do. The present work just sketches a first step towards understanding the statistics. Our most important conclusion is that when random textures increase in density their statistics, especially modulation power, tends towards that of noise. However in the case of quasi-periodic textures, modulation power could be used for some sort of pitch-shifting purposes. For further investi-

gations we would need to refine our methodology and start by identifying higher-level perceptual parameters that apply to all textures. But even if we were to find such a set of parameters, the definition of each parameter could vary from one texture type to another. The density of a wind sound is probably something very different to the density of a rain sound, and could yield completely different statistical behavior.
This work mainly provided a analysis of what McDermott's model could do, discussed its limitations and made the link with a the theoretical definition of sound textures. We also presented a first attempt to describe how higher-level parameters, here through the example of density, could be coded in a statistical framework. This latter investigations would need to be refined in further research work.

References

Athineos, M., and Ellis, P. (2003). Sound Texture Modeling with Linear Prediction in both Time and Frequency Domains. Proc. ICASSP-03.

Dubnov, S., Bar-Joseph, Z., El-Yaniv, R., Lischinski, D., and Werman, M. (2002). Synthesis of audio sound textures by learning and resampling of wavelet trees. IEEE Computer Graphics and Applications

Erdreich, J. (1986). A distribution based definition of impulsive noise. J. Acoust. Soc. Am. Volume 79, Issue 4, pp. 990-998.

Hoskinson, H., and Pai, D. (2001). Manipulation and Resynthesis with Natural Grains. Proceedings of the International Computer Music Conference (ICMC), pages 338–341.

Lu, L., Wenyin, L., and Zhang (2004). Audio Textures: Theory and Applications. IEEE Transactions on Speech and Audio Processing, Vol. 12, No. 2.

McDermott, J.H., and Simoncelli, E.P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery, Evidence from Sound Synthesis. Neuron.

McDermott, J.H., Oxenham, A.J., and Simoncelli, E.P. (2009) Sound Texture Synthesis via Filter Statistics. IEEE Workshop on Application of Signal Processing to Audio and Acoustics.

McDermott, J.H., Schemitsch, M., and Simoncelli, E.P. (2013). Summary Statistics in Auditory Perception. Nature America.

Misra, A., Cook, P.R., and Wang, G. (2006). A New Paradigm for Sound Design. Proc. of the 9th Int. Conference on Digital Audio Effects (DAFx-06).

Mölhmann, D. (2011). A Parametric Sound Object Model for Sound Texture Synthesis. Phd Thesis, University of Bremen.

Portilla, J., and Simoncelli, E.P. (1999). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. International Journal of Computer Vision.

Saint-Arnaud, N., and Popat, K. (1995). Analysis and synthesis of sound textures. Readings in Computational Auditory Scene Analysis, pp. 125–131.

Schaeffer, P. (1966). Traité des Objets Musicaux, p. 454.

Schwarz, D., Beller, G., Verbrugghe, B., and Britton, S. (2006). Real-time Corpus Based Concatenative Synthesis with Catart. Proceedings of the 9th. Int. Conference on Digital Audio Effects (DAFx-06).

Strobl, G., Eckel, G., and Rocchesso, D. (2006). Sound Texture Modeling : A Survey. SMC Conference.

Zhu, X., and Wyse, L. (2004). Sound texture modeling and time-frequency LPC. Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx), volume 4.

7 Appendix 1: Artificial Sound Textures

Analysing only recorded sound textures brings some limitations because from one example to another we cannot control exactly how parameters are varying. For example two rain sounds are never produced in the exactly same conditions and when we choose two samples of different densities there is actually a number of other parameters that could change, like the type of ground - grass, concrete, etc - or the point from where the sound is recorded, the type of microphone used...
In order to gain more control over what we are studying we choose to synthesize artificial textures.

7.1 Random noise bursts streams

These artificial texture are inspired by Saint-Arnaud's two-level description of sound textures. The individual events were generated by inverse Fourier transform. Frequency regions - which bandwidth and center frequency were randomized - were convolved with the Fourier transform of a Hanning window - which time width was also randomized - thus yielding Hanning shaped time events. Amplitude of each event could also be randomized.
The events were then put in a stream. The spacing between two events was randomized, but constrained on one hand by the desired density (defined as the number of events per second) and on the other hand by a "periodicity" parameter . A "periodicity" of one meant a completely periodic texture, density thus being frequency in that case.
The generated textures had a bandwidth between 200 Hz and 10000 Hz, a time width between 20 ms and 50 ms and a periodicity of 0.5. All random distributions were uniform.

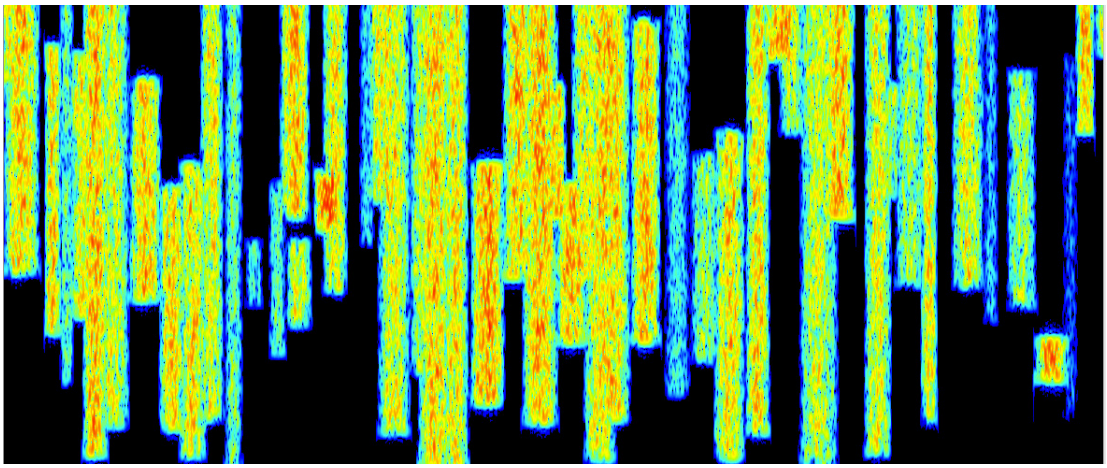


Figure 19: Example artificial texture. Density is around 10 events per second.

Periodic noise bursts streams

Periodic noise bursts streams were produce with the same method as presented previously. We removed all randomization in time and frequency and made the events broadband. Density thus here corresponds to the rate of the events.

7.2 Artificial helicopter

The synthesis of helicopter sound was inspired by a recorded sample of helicopter. The original helicopter sound is mainly composed of low frequency noise bursts with static background noise. We synthesized artificial helicopter sounds by generating a spectrally matched noise based on the original helicopter sound and then by adding low frequency noise bursts generated with the previous method. The spectrally matched noisee is basically white noise which is filtered so that it has the same time-average spectrum as the original sound. Figure [20] presents original version and artificial.

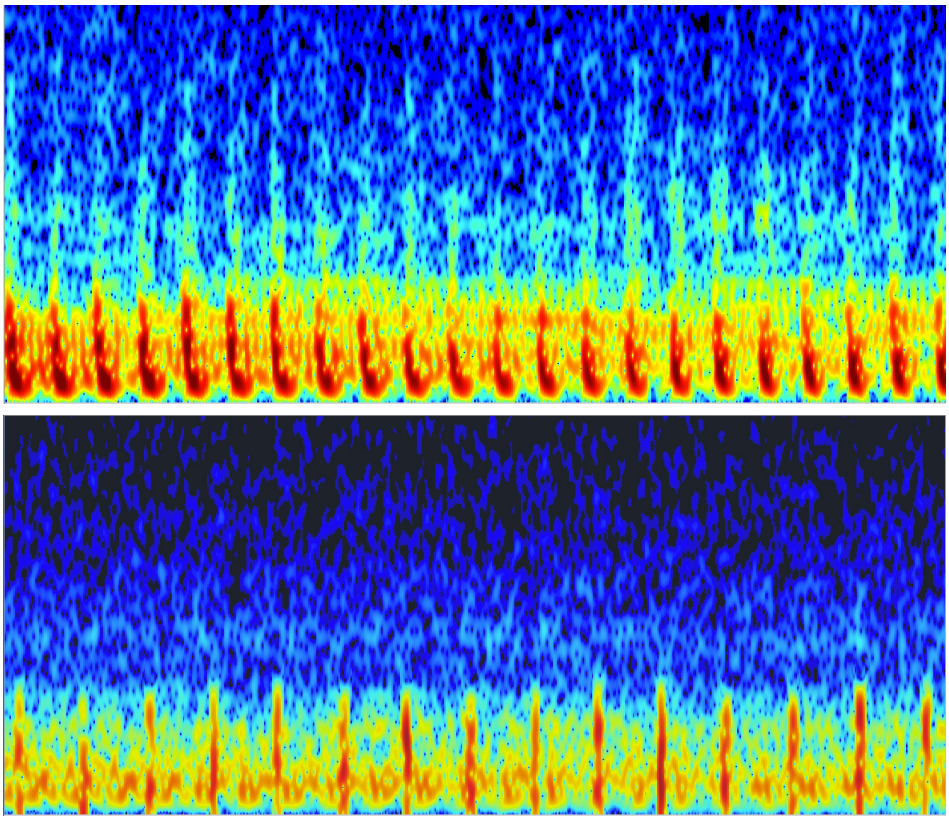


Figure 20: Example of re-synthesis of a helicopter sound. Top is original, bottom is synthesized.