TIME SCALE MODIFICATION OF NOISES USING A SPECTRAL AND STATISTICAL MODEL

Pierre Hanna

SCRIME - LaBRI
Université de Bordeaux 1
F-33405 Talence Cedex, France
hanna@labri.fr

Myriam Desainte-Catherine

SCRIME
Université de Bordeaux 1
F-33405 Talence Cedex, France
myriam@labri.fr

ABSTRACT

Some natural sounds, such as speech parts can essentially be considered as noises. For instance, models suppose noisy parts of sounds as weak parts and apply basic approximations. But transformations such as time stretching doesn't preserve the noisy characteristics of sounds. Moreover, we show that those transformations introduce artificial intensity variations. In this paper we propose a spectral model for noise modeling which takes into account the statistical properties of such sounds. The analysis is based on the classical spectral models. The synthesis consists of randomly defining sinusoidal components. These components are then added using adapted overlap-add method to keep statistical moments constant. Time scaling operations using this approach are described. Experiments on artificial sounds (filtered white noises) as well as natural sounds such as consonants and whispered vowels, show impressive enhancement in quality. Infinite time stretching transformations of such noises can be perfectly performed.

1. INTRODUCTION

Generalized spectral model separates the analyzed sound into deterministic and stochastic parts [1]. The deterministic part is a sum of slow-evolving sinusoids. The stochastic part, called the residual, is obtained by subtracting the magnitude spectra of the deterministic signal from that of the original. This stochastic part is composed of all the signal components that haven't been retained by the phase vocoder. Its presence is very important in order to improve the realism of the synthesized sounds. This part can be considered as random signal with specific statistical properties. Time scaling transformations of the deterministic part have been well studied and have very good performance. But concerning the stochastic part, the usual spectral representation is dedicated to the *residual* part of a signal [2], which means that its proportion in the original signal is very weak (breath noise in wind instrument as flute or saxophone). Hence basic methods suffer from limitations for time modifications. Furthermore, some natural sounds are composed of a major part of stochastic signals (consonants, whispered voices, drums, ...) and such approximations imply audible limitations.

2. BACKGROUND

The existing models to analyze, transform and synthesize noisy sounds are temporal or spectral models. Temporal models generate noises by randomly drawing samples using a standard distribution (uniform, normal, etc...). Then, they may be filtered (subtractive synthesis). The main temporal models use linear predictive coding (LPC) to color white noise source. These approaches are common in speech research but are less intuitive for spectral transformations and are less flexible: the smoother the spectral envelope is defined, the more complicated is the computation of the coefficients.

Moreover we are particularly interested in spectral models homogeneous with other spectral models for harmonic sounds (for example [3]). Many works concern stochastic part modeling. Some of them [1, 4] propose to represent this part with overlapped amplitude spectra. These spectra are approximated with simple line-segments. The synthesis of this component consists in generating white noise colored with analyzed amplitude spectra and random phase spectra (uniform distribution), by performing an inverse Fast Fourier transform (IFFT). We show in section 6 that transformations such as time scaling imply some variations of the statistical properties of the original sounds, and thus audible artifacts.

A residual model related to the properties of the auditory system is proposed in [2]. The noisy part of any sound is represented by the time-varying energy in each equivalent rectangular band (ERB). However, because of such approximations, this model may be applied only to sounds with weak stochastic parts.

In this paper we present analysis and synthesis methods that are based on the stochastic part model, but which can be applied to noisy sounds and which allow perfect time scaling transformations.

3. SPECTRAL AND STATISTICAL MODEL

Sounds (sample rate F_e) are considered as random processes X. They are modeled as a sum of sinusoidal components. The i^{th} component has frequency f_i which is a random variable with fixed amplitude a_i and uniformly distributed phase ϕ_i:

$$X_k = \sum_{i=0}^{N} a_i \sin(2\pi f_i \frac{k}{F_e} + \phi_i) \qquad (1)$$

where the frequencies f_i are distributed in a band whose width is ΔF (Hz).

The draw of the frequency values are described in [5]. M bins ($M \leq N$) equally divide the frequency bandwidth. In each successive frame, N frequency values are chosen into these bins according to a uniform distribution. Therefore the probability that

two successive frames have exactly the same amplitude spectrum, is very low.

In the following, the main synthesis parameter is the size W_s of the synthesis window. The number of sinusoidal components and the number of bins are set as half of the synthesis window size W_s.

$$M = N = \frac{W_s}{2} \tag{2}$$

This value corresponds to the maximum value of the spectral density. Choosing more sinusoidal components would not have any perceptual influence for the synthesized sound [5]. This choice is also implicitly done in the inverse-Fourier transform (SMS model [1]). For a $2k$ samples long window, the stochastic part is represented by the sum of k fixed sinusoids.

It is useful to note that this number of sinusoidal components is perceptually relevant and that controlling this parameter allows modifications of the spectral density [6], which are not possible using other representations.

4. ANALYSIS

The analysis part is very simple and similar to traditional ones (see figure 1). It starts by computing amplitude spectra of sound under study using the Fourier transform. Although our method concerns noisy sounds (more generally random signals), more precision about the time variations is needed. Hence we take successive analysis windows x_l that are overlapping:

$$x[n] = \sum_{l=0}^{L-1} x_l[n - lH] \tag{3}$$

where H is the hop size (or the time advance) and l is the frame number ($l \in 0, .., L-1$).

Each temporal window of the original signal s is multiplied by a weighting window w:

$$x_l[n] = s_l[n]w[n] \quad n \in [0, 1, .., W_s - 1] \tag{4}$$

The choice of the length of the window deals with the usual trade-off of time versus frequency precision. Choosing a short window leads to smooth spectra whereas a long one sets the assumption of stationarity. Spectra are interpolated using zero-padding. Experiments show that 4096 samples are generally sufficient to have a good time and frequency precision. The type of the analysis window can also be discussed. With the usual methods (SMS, STN), the overlap-add synthesis imposes the use of the same window in the analysis part and in the synthesis part. However, any type of window can be used with our synthesis method.

The choice of the hop size H depends on the efficiency of the analysis (and sometimes, synthesis) process. A small value improves the accuracy in time domain whereas a large value decreases the computation time. During our experiments, we use hop size equal to half or one-fourth of the analysis window size.

Fig. 1. *Analysis block diagram*

5. SYNTHESIS

We describe here the synthesis process (see figure 2). As previously seen in section 3, the number of frequency components is linked to the synthesis window size W_s. The synthesis part starts by defining the bins B_i ($i \in \{0, .., N-1\}$) from this size and the bandwidth $\Delta F = 22050$Hz:

$$B_i = [i\frac{\Delta F}{N}; (i+1)\frac{\Delta F}{N}[\tag{5}$$

Then, for each synthesis frame l ($l \in \{0, .., L-1\}$), one frequency value is randomly chosen into each bin according to a uniform distribution:

$$f_i^l = (i + \texttt{rand}(1.0))\frac{\Delta F}{N} \tag{6}$$

where `rand` represents the *classical* random function which returns a pseudo-random real between 0 and the parameter of the function.

The phase values of each sinusoidal component are randomly chosen according a uniform distribution between 0 and 2π:

$$\phi_i^l = \texttt{rand}(2\pi) \tag{7}$$

The amplitude of each sinusoidal component is calculated from the amplitude spectrum and the frequency value. Since this value is randomly chosen, its amplitude may be interpolated. Many methods can be used (for example [1]) which are not described here. The simplest one is to interpolate the spectra during the analysis process using zero-padding.

Once the frequency, amplitude and phase values are calculated, temporal samples of each frame are generated with additive synthesis. An efficient algorithm is presented in [3].

The resulting temporal signal does not taper to 0 at the boundaries of each window because of the random values of phase. Hence synthesized windows are overlapped and added to avoid audible clicks. This OLA synthesis needs to multiply each frame by a weighting window. It is important to note that the type and the length of the synthesis window are independent of that of the analysis window.

5.1. Preserving the statistical moment

A noisy sound can be seen as a random signal: s_l ($l \in \{0, .., L-1\}$) are realizations of the same random variable X. An accurate synthesis method needs to preserve the statistical moments of the original signal, even if transformations such as time or frequency scaling are performed. Otherwise, audible artifacts can be heard.

Indeed one can show that the variance $v = \sigma^2$ is directly linked to the perceived intensity of the sound, for example, for a uniform or a Gaussian white noise:

$$A_{RMS} = \sigma \tag{8}$$

Let X' be the random variable associated with the synthesized signal $x'[n]$. Then the mean of X' equals the mean of X. Let $V(X')$ be the variance of X'. If an analysis is performed but no transformation is done, s_l are not independent and we verify $V(X') = V(X)$. We consider here the case when there are transformations or amplitude spectra approximations. Since s_l are then independent and are realizations of the same random variable X, one can write:

$$V(X') = V(X)\sum_{l=0}^{L-1} w^2(n - lH) \tag{9}$$

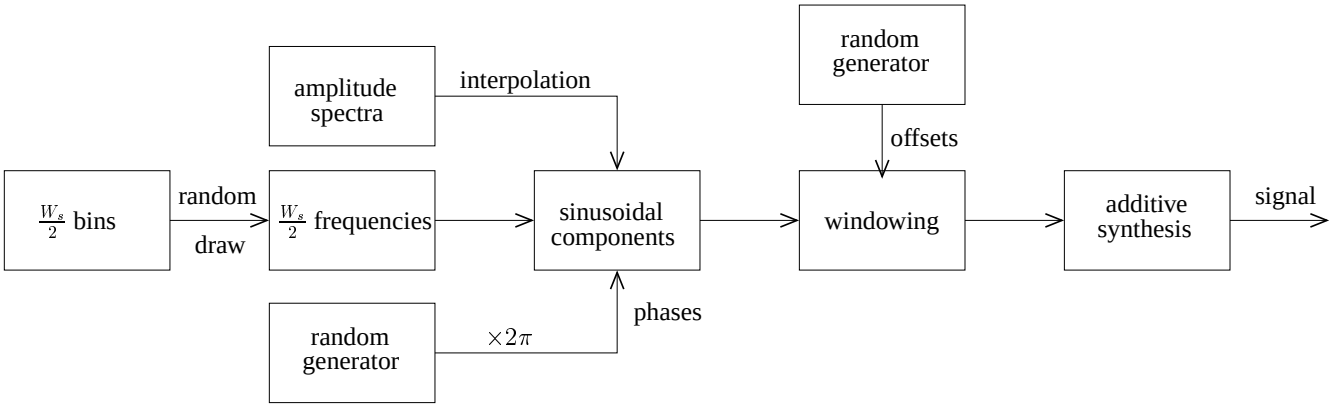

Fig. 2. *Synthesis block diagram*

Assuming that $V(X')$ and $V(X)$ are equal, this calculus leads to the condition about w, and for all $n \in \mathbb{N}$:

$$\sum_{l=0}^{L-1} w^2(n - lH) = 1 \tag{10}$$

This equality is only verified for some window types (sinusoidal for example), but not the usual ones (Hann, Bartlett, etc...) used in [1, 4]. In these cases the intensity of the synthesized sound vary periodically in time.

Figure 3 shows the variations of the variance as a function of time with $N = 16384$ and $H = \frac{N}{2}$.

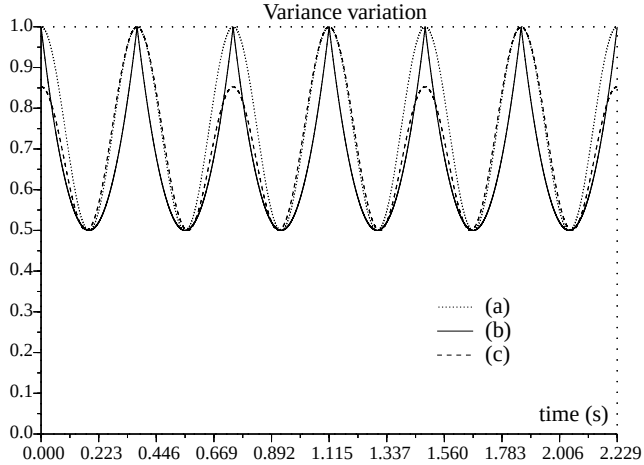

Fig. 3. *Variations of the variance of the signal with (a) Hann window (b) Bartlett (triangular) window (c) Hamming window.*

5.2. Synthesis methods

We have proposed two ways [7] to avoid these intensity variations which imply audible artifacts. The first one consists in using a sinusoidal window as weighting window which verifies the condition 10. This method is the simplest and improves the quality of the synthesized sound. The second one randomly shifts the starting time of each sinusoidal components which are separately multiplied by a weighting window. In this case any type of window can be used. Both methods lead to perceptually good results. Discussions about advantages and drawbacks of each method can be found in [7].

6. TIME SCALING

Time scaling is an useful but difficult transformation. This process must essentially preserve the noisy characteristics of the original sounds. By using our proposed analysis and synthesis methods for the stochastic part of the original sound, transformations such as time or frequency scaling can be perfectly performed. Usual models [1, 4] stretch sounds by modifying the hop size or increasing the synthesis window size. We show in the previous section that such modification changes the statistical properties of the sound and provokes audible intensity variations. Experiments confirm this theory.

Moreover, these two possibilities imply other drawbacks. In one hand increasing this hop size obviously limits the stretching factor. The technique we propose doesn't impose any limit. On the other hand extending the size of the synthesis window needs spectra interpolation. If the size is too large compared to the analysis window size, the spectra approximations degrade the output signal. Furthermore, this method imposes more CPU consumption.

Our synthesis method allows different possibilities to stretch sounds. We choose to preserve the hop size in order to keep the number of sinusoidal components constant, because the spectral density of sounds [6] is perceptually relevant. We focus on time stretching modification, but slowing down sounds is based on the same principle. In our approach, the simplest and the most efficient technique concerns dilatation factor α that is multiple of the ratio between the synthesis window size W_s^s and the analysis window size W_s^a. In this case, the number of synthesis windows is simply increased. Obviously new sinusoidal components are randomly determined in each new frame. This statistical characteristic makes this choice possible at the contrary of IFFT-based methods which impose constant amplitude spectra and lead to *metallic* sounds.

If α doesn't verify this condition, it can be written as $\alpha = k\frac{W_s^s}{W_s^a} + \alpha_0$, where $k \in \mathbb{N}$ and $\alpha_0 < \frac{W_s^s}{W_s^a}$ and therefore, the size of the synthesis window is modified: $W_s^{s'} = (1 + \alpha_0)W_s^s$.

In our approach, as previously seen, increasing the synthesis window size without degrading the signal imply increasing the

number of sinusoidal components: $N' = \frac{W_s^{s'}}{2}$. This operation may impose an interpolation of the amplitude spectra, but at the difference of the others methods, this interpolation is always limited according to the ratio of the analysis and synthesis window size. Hence, during our experiments, we prefer choosing a large analysis window size compared to synthesis window size.

Generating new windows doesn't consume more time because the same number of sinusoidal components is synthesized per time unity. At the opposite, increasing the synthesis window size requires more CPU consumption because the number of oscillators also increases. Since the increase of the number of synthesis window is privileged against its size, such operation is restricted to small variations ($\alpha_0 < \frac{W_s^s}{W_s^a}$) and thus imposes small CPU consumption variations.

7. APPLICATIONS

This model is implemented as a free library and a jmax [8] object. They are available on the website of the SCRIME [9]. We have performed several experiments concerning different types of sounds. In this paper we don't focus on the detection and analysis of fast attacks (or transients). Our tests were restricted to sounds without transients. The first experiments concern synthetic noises (filtered white noises). Measures of statistical moments confirm that our method doesn't introduce the same artificial intensity fluctuations as the classical methods do. Other experiments have been done with speech sounds. Unvoiced consonants and whispered sentences are perfectly synthesized and time extended with the same perceptual quality. The size of the analysis window was fixed to 4096 samples and that of the synthesis window to 1024. These choices correspond to the best quality for time stretching.

Sound examples are available at `http://www.labri.fr/Perso/~hanna/`.

8. CONCLUSION AND FUTURE WORK

The original spectral and statistical model we propose permits high quality synthesis of noisy sounds and allows infinite time stretching without degrading the original sound. In our process, the analysis and synthesis windows are totally independent (type, length, hop size). Moreover, this model is homogeneous with other harmonic models restricted to deterministic parts and may greatly complement it to make sounds more realistic. In the future we will improve this model in order to take into account transients and to preserve it during time transformations.

9. ACKNOWLEDGMENTS

This research was carried out in the context of the SCRIME[1] project which is funded by the DMDTS of the French Culture Ministry, the Aquitaine Regional Council, the General Council of the Gironde Department and IDDAC of the Gironde Department.

SCRIME project is the result of a cooperation convention between the Conservatoire National de Région of Bordeaux, ENSEIRB (school of electronic and computer scientist engineers) and the University of Sciences of Bordeaux. It is composed of electroacoustic music composers and scientific researchers. It is managed by the LaBRI (laboratory of computer science of Bordeaux). Its main missions are research and creation, diffusion and pedagogy thus extending its influence.

10. REFERENCES

[1] X. Serra and J. Smith, "Spectral modeling synthesis: a sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[2] M. Goodwin, "Residual modeling in music analysis-synthesis," *Proceedings of the IEEE International Conference On Acoustics, Speech, and Signal Processing (ICASSP'96), Atlanta, GA*, pp. pp. 1005–1008, 1996.

[3] S. Marchand, *Sound models for computer music: analysis, transformation, synthesis of musical sound*, Ph.D Thesis, LaBRI, Université Bordeaux I, 2000.

[4] T.S. Verma and T.H.Y. Meng, "Time scale modification using a sines+transient+noise signal model," *Proceedings of the Digital Audio Effects Workshop (DAFX'98, Barcelona)*, pp. 49–52, 1998.

[5] P. Hanna and M. Desainte-Catherine, "Real-time noise synthesis with control of the spectral density," *Proceedings of the Digital Audio Effects Workshop (DAFX'02, Hamburg, Germany)*, pp. 151–156, 2002.

[6] W.M. Hartmann, S. McAdams, A. Gerzso, and P. Boulez, "Discrimination of spectral density," *Journal of Acoustical Society of America*, vol. 79, no. 6, pp. 1915–1925, 1986.

[7] P. Hanna and M. Desainte-Catherine, "Adapting the overlap-add method to the synthesis of noise," *Proceedings of the Digital Audio Effects Workshop (DAFX'02, Hamburg, Germany)*, pp. 101–104, 2002.

[8] "jmax," `http://www.ircam.fr/jmax`.

[9] "Scrime," `http://www.scrime.u-bordeaux.fr`.

[1] Studio de Création et de Recherche en Informatique et Musique électroacoustique