ON THE MODELING OF SOUND TEXTURES BASED ON THE STFT REPRESENTATION

Wei-Hsiang Liao
UMR STMS IRCAM - CNRS - UPMC
Paris, France
wliao@ircam.fr

Axel Roebel
UMR STMS IRCAM - CNRS - UPMC
Paris, France
roebel@ircam.fr

Alvin W.Y. Su
SCREAM Lab, Dept. of CSIE
Nat. Cheng-Kung Univ.
Tainan, Taiwan
alvinsu@mail.ncku.edu.tw

ABSTRACT

Sound textures are often noisy and chaotic. The processing of these sounds must be based on the statistics of its corresponding time-frequency representation. In order to transform sound textures with existing mechanisms, a statistical model based on the STFT representation is favored. In this article, the relation between statistics of a sound texture and its time-frequency representation is explored. We proposed an algorithm to extract and modify the statistical properties of a sound texture based on its STFT representation. It allows us to extract the statistical model of a sound texture and resynthesise the sound texture after modifications have been made. It could also be used to generate new samples of the sound texture from a given sample. The results of the experiment show that the algorithm is capable of generating high quality sounds from an extracted model. This result could serve as a basis for transformations like morphing or high-level control of sound textures.

1. INTRODUCTION

Environmental sounds such as wind, rain or crowd noises are common in our daily life. In contrast to instrumental sounds, these sounds are often inharmonic, chaotic and possess stronger noise components. They are often known as sound textures.

Recently, sound texture processing has become more important in the entertainment industry and various other areas. High quality sound texture transformations are essential for work in these fields. State-of-the-art algorithms are capable of achieving high-quality transformations with sinusoidal sounds[1] [2] [3] [4], while preserving the envelope and naturality of transients [5] [6] [7]. However, these algorithms yield less satisfying results when applied to sound textures. The underlying reason is that sound textures do not fit the common assumptions of most state-of-the-art algorithms, such as harmonicity and sinusoidality.

Sound textures are often driven by a stochastic process. They usually exhibit some type of structure over a moderate time period. Julesz's conjecture[8] and [9] suggests that the structure can usually be described by a series of statistical properties. Moreover, in [10] and [11], it is stated that humans distinguish between sound textures based on their statistical properties. Based on these assumptions, it is possible to establish a parametrized model which describes a sound texture using a set of statistical properties. This model will allow the analysis, synthesis and transformation of sound textures.

There are several previous works on sound textures. Saint-Arnaud[12] proposed an analysis/synthesis scheme based on modeling the atomic events in sound textures. In [13], stochastic noises are synthesized with randomized sinusoids. Schwarz[14] proposed a descriptor-driven, corpus-based approach to synthesize sound textures. In addition, several parametric models have been proposed for image textures. For example, Portilla[9] proposed a texture model based on the statistics of wavelet coefficients. Bruna[15] proposed a new wavelet transform which provides a better view of textures while capturing high-order statistics. For sound textures, McDermott[10] proposed a parametrized model, adapted from [9], which characterizes target sound textures with high order statistics and correlation between subband envelopes. In his article, statistical properties are applied to Gaussian noises to generate different sound textures.

A common approach of signal processing is to use the STFT (Short Time Fourier Transform) representation. The idea is to use STFT to obtain the time-frequency representation of the signal and apply the transformation. However, conventional transformation techniques, such as phase vocoder, cannot be applied to sound textures[16]. To apply transformations to a sound texture over its STFT representation, a statistical model based on the signal's time-frequency representation is required. Also, establishing a model based on the STFT representation can allow us to utilize existing efficient implementations of DFT.

In this article, we would like to propose a model which can be used to analyse and synthesise sound textures. It's based on the statistics of time-frequency representations of sound textures. From the model, one could generate new samples of the input texture. It could serve as a basis for high-level transformations of sound textures. An experiment has been conducted to investigate the quality of sound generated from the model.

The paper is organized as follows: The first section describes the present state of sound texture processing and the importance of establishing a statistical model. The second section discusses perceptually important statistics of sound textures and the set of properties to be used in the model. The third section describes the process of modeling a sample texture and resynthesising it with

the resulting model. The fourth section is an experiment of texture generation with the model. The fifth section are the conclusion and future works.

2. SOUND TEXTURE STATISTICS

Mcdermott's work[10] suggests that a proper model of a sound texture is composed of the statistical properties of individual subband signals. These subband signals can be formed from the time-frequency representation of the original signal. Furthermore, from his experiment, we know that the perceptually important statistical properties can be summarized into three categories : *Subband Moments*, *Temporal Correlation* and *Spectral Correlation*.

Subband Moments

Subband moments are the statistical moments of the spectral coefficients of subbands. They describe the shape of the histogram of a subband signal.

Temporal Correlation

Temporal correlation refers to the autocorrelation of a subband signal. It can be characterized by the autocorrelation function.

Spectral Correlation

Spectral correlation means the cross-correlation between different subband signals.

While the autocorrelation function and cross-correlation characterize the horizontal and vertical relationships in the time-frequency representation, slant relationships should also be considered. This can be characterized by the delayed cross-correlation, which is the cross-correlation function. As such, the three categories could be summarized in two, where moments characterize the property of one subband signal, and correlation functions characterize the relationships between subband signals in all directions.

Here, the proposed model uses STFT as the underlying time-frequency representation. Each bin of the STFT representation is treated as a single subband, and the coefficients of each subband form the subband signal. Since humans mainly perceive envelope statistics[10], the statistics of the subband signals and their envelopes are both included. In order to establish a model which contains all the statistics from the categories described above, the following statistics are included: the first four subband moments, the auto-correlation function of each subband, and the cross-correlation function of each pair of subbands.

3. MODELING SOUND TEXTURES

In this section we describe the proposed modeling mechanism. The first subsection describes the process by which we extracts the model from a given sound texture. The second subsection describes how to apply the model to a STFT representation and the steps to follow to resynthesize sounds.

3.1. Model Extraction

Assume a given input signal x, along with a window function w, hop size h and the size of Fourier transform N. Applying STFT to the signal will obtain the analysis frames. The analysis frame X_l which is centered at time point l is:

$$X_l(k) = \sum_t x(t)w(t-lh)e^{\frac{-j2\pi tk}{N}} \tag{1}$$

These analysis frames X_l form the STFT representation. Thus, the subband signals are formed as:

$$S_k(n) = X_n(k)e^{\frac{j2\pi lhk}{N}} \tag{2}$$

In (2), a demodulation term is applied to remove the carrier frequency in each subband. The carrier frequency is introduced by the STFT and will affect the evaluation of correlation functions. Next, we extract the envelope E of the subband signal, which can be done by any envelope extraction algorithm. One could also use the amplitude of the subband signal for simplicity.

$$E_k(n) = envelope(S_k(n)) \tag{3}$$

At last, statistics are evaluated for every S_k and E_k. Here, only the moments of the envelopes E_k are evaluated. This is because currently we have not yet found a proper way to define complex moments for the proposed model. Finally, the extracted model Φ of the input texture can be written as:

$$\Phi \equiv \{A_S, C_{SS}, M_E, A_E, C_{EE}\} \tag{4}$$

$$A_{S_k}(l) = \{\sum_t S_k(t)S_k(t+l)\} \tag{5}$$

$$C_{S_xS_y}(l) = \{\sum_t S_x(t)S_y(t+l)\} \tag{6}$$

$$M_{E_k} = \{\mu_1(E_k), \mu_2(E_k), \eta(E_k), \kappa(E_k)\} \tag{7}$$

$$A_{E_k}(l) = \{\sum_t E_k(t)E_k(t+l)\} \tag{8}$$

$$C_{E_xE_y}(l) = \{\sum_t E_x(t)E_y(t+l)\} \tag{9}$$

In (4), A denotes the autocorrelation function, C denotes the cross-correlation function and M denotes the first four moments, namely mean, variance, skewness and kurtosis. Where mean and variance are central moments, skewness and kurtosis are normalized central moments. After the model has been extracted, certain transformations could be applied to it. For example, time-stretching is equal to stretching the autocorrelation and cross-correlation functions.

3.2. Texture Resynthesis

Sound textures can be generated from the model extracted in the previous subsection. The idea is to apply the statistical properties to a randomized STFT representation. Usually, it requires immense iterative computation to estimate the output signal. However, this can be avoided by utilizing the relationship between the correlation and the spectrum.

3.2.1. Applying autocorrelation

To apply the autocorrelation functions, we could use the Wiener-Khinchin theorem. It suggests that the amplitude spectrum of a signal is also the spectrum of its auto-correlation function.

$$\mathcal{F}(A_s) = \mathcal{F}(s(t) * \overline{s}(-t)) = |\mathcal{F}(s)|^2 \tag{10}$$

That is, we can assign any autocorrelation function to a subband signal by constraining its amplitude spectrum. To avoid circular aliasing, a Fourier transform with twice the length of the autocorrelation function is used when calculating the spectrum of the subband signal. This introduces a windowing effect to the signal. We use Griffin Lim's[17] technique to re-establish the window shape after the inversion.

3.2.2. Applying cross-correlation

The cross-correlation functions can be dealt with in a similar fashion. Here we introduce an extension of (10), the cross-correlation theorem:

$$\begin{aligned} \mathcal{F}(C_{x,y}) &= \mathcal{F}(x(t) * \overline{y}(-t)) \\ &= \mathcal{F}(x(t))\mathcal{F}(\overline{y}(t)) \\ &= |X(k)||Y(k)|e^{j(\theta(X(k))-\theta(Y(k)))} \end{aligned} \quad (11)$$

Equation (11) can be rewritten as:

$$\theta(\frac{\mathcal{F}(C_{x,y})}{|X(k)||Y(k)|}) = \theta(X(k)) - \theta(Y(k)) \quad (12)$$

Since we already constrained the amplitude spectrum by (10), we could control the cross-correlation function between x and y by adjusting their phase differences without affecting the autocorrelation functions. To establish full cross-correlation across all subband signals, choose one subband signal as the starting point, initialize phases for the spectrum of the chosen subband signal, assign phases to neighboring subbbands with the phase differences extracted from the original input texture, as in (12). Here we choose the zeroth bin ,S_0, as the starting point. Since S_0 is real, the phase of it's spectrum must be Hermitian, that is, $\theta(\mathcal{F}_{S_0}(k)) = -\theta(\mathcal{F}_{S_0}(-k))$.

3.2.3. Applying Moments

The moments only apply to envelopes. Since mean and variance are already contained in the autocorrelation function, the remaining moments are the skewness and kurtosis. A gradient descent algorithm is used to apply the skewness and kurtosis, the detail of which can be found in [9]. The following formula is used to apply the moments to a signal x:

$$x' = x + \gamma \frac{\partial f(x)}{\partial x} \quad (13)$$

The target signal x' can be obtained by solving γ with respect to $f(x') = t$. If there is no real root for γ, a real-value approximation which is closest to the target value will be chosen. In (13), f is either skewness or kurtosis, where $\frac{\partial f(x)}{\partial x}$ is the derivative of the moment function respected to x, and t is the target value of the desired moment. For skewness η and kurtosis κ, their derivatives with respect to x are:

$$\frac{\partial \eta(x)}{\partial x} \equiv x \circ x - \mu_2^{1/2}(x)\eta(x) - \mu_2(x) \quad (14)$$

$$\frac{\partial \kappa(x)}{\partial x} \equiv x \circ x \circ x - \mu_2(x)\kappa(x)x - \mu_3(x) \quad (15)$$

With the gradient descent algorithm, we could apply the moments to the subband signals. Unfortunately, the gradient descent algorithm would change the correlation functions of the subband signals, so, an iterative approach is used to generate the target STFT representation.

The whole resynthesis step is depicted in fig.1. It begins with a randomly initialized STFT representation. Next, the STFT representation is decomposed into the envelope part and the remaining part. Envelope statistics $\{A_E, C_{EE}, M_E\}$ are applied by the methods described in previous sections. Then, reconstruct the STFT representation by combining the modified envelope and the remaining part, and apply the subband signal statistics $\{A_S, C_{SS}\}$ accordingly. If the iteration has converged or reached maximum iteration number, output the STFT representation, otherwise, repeat the iteration process. At last, apply the remodulation, inverse STFT and overlap-add to generate the resynthesized signal $\tilde{x}$:

$$\begin{aligned} x'(n) &= \sum_k S_k(l) e^{\frac{j2\pi k(n-lh)}{N}} \\ \tilde{x}(n) &= \frac{\sum_n w(n-lh)x'(n)}{\sum w^2(n-lh)} \end{aligned} \quad (16)$$

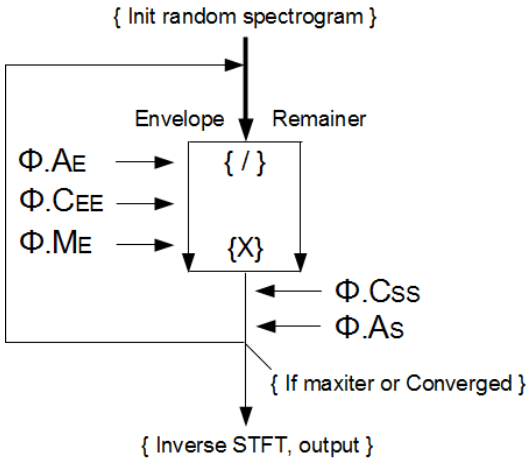
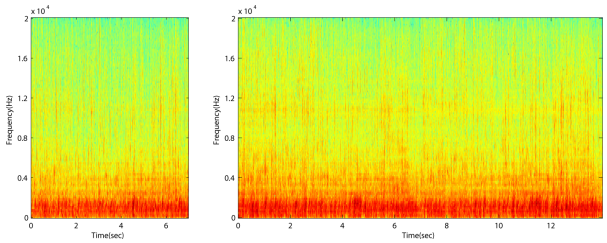


Figure 1: *The flow diagram of the resynthesis procedure*

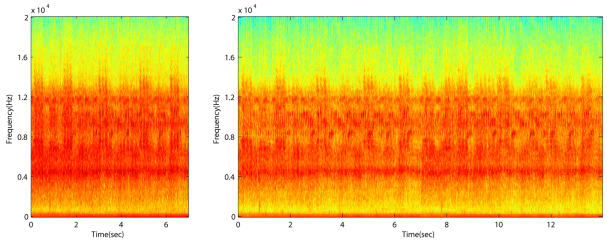
4. EXPERIMENT

To demonstrate the texture sound generated from the model, we used some texture samples from [10]. Statistical models were extracted from the samples of sound textures, then new samples were regenerated from the model. To generate a texture longer than the original, the model is used twice and connected with overlap-and-add. The discontinuation between generated blocks will be handled in the future. The detailed configuration is listed below:

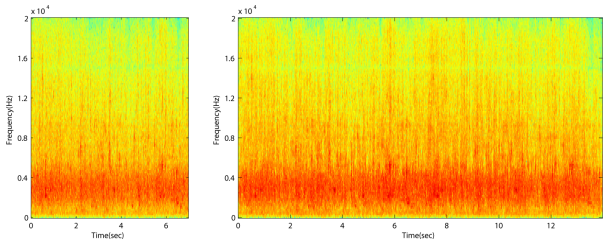
- number of samples : 6
- original signal : 16bit mono wav, 5-7 secs
- sampling rate : 20khz
- hop size of input frame : 16 samples
- size of Fourier transform : 256, hanning window
- generated sample length : 10-14 secs (2 times long)



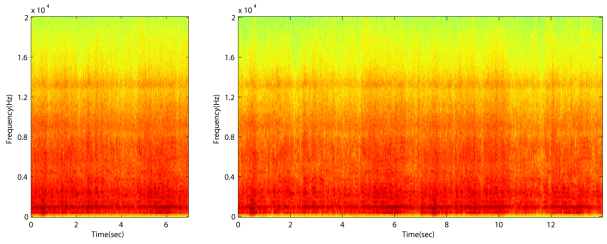
(a) *Bubbling Water*



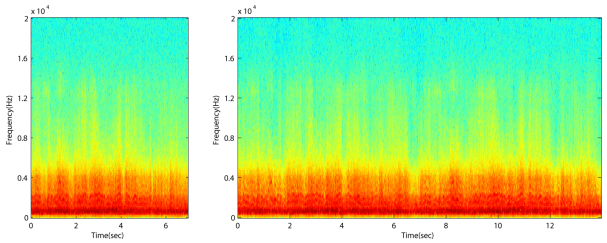
(b) *Insects In A Swamp*



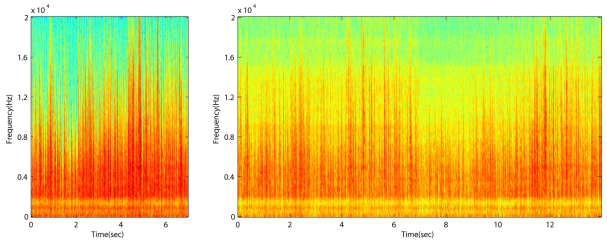
(c) *Heavy Rain Falling And Dripping*



(d) *Bees*



(e) *Wind Whistling*



(f) *Rustling Papers*

Figure 2: *The spectrograms of original(left) and generated(right) textures.*

The generated sound samples can be found in [18]. From the quality of generated sounds, we could see the proposed model is capable of capturing most of the perceptually important features. However, the iterative process is not guaranteed to converge to a good solution, as the case of fig.2f. It may require more constraints to improve the quality of the click sounds.

5. CONCLUSION

In this paper, we proposed a parametric statistical model based on the STFT representation. It can be used to analyse and synthesise sound textures and generate high quality sounds. Since it's based on the STFT, it could benefit from many existing related implementations. Possible applications of this model are time-stretching, texture morphing and providing high-level control of sound textures. Integrating the model with the phase vocoder could also be a future goal. The future works seek a way to apply all the statistics at once (eliminating the iterative process), connect the model with semantic controls and investigate an efficient way to apply only specific parts of correlation functions.

6. ACKNOWLEDGMENTS

This research is funded by the French National Research Agency(ANR) project PHYSIS. Thanks to Alex Baker for proof-reading the manuscript.

7. REFERENCES

[1] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[2] M-H. Serra, "Musical signal processing, chapter introducing the phase vocoder," in *Studies on New Music Research. Swets & Zeitlinger*, 1997, pp. 31–91.

[3] J. Laroche and M. Dolson, "New phase-vocoder techniques for real-time pitch shifting, chorusing, harmonizing and other exotic audio modifications," *Journal of the AES*, vol. 47, no. 11, pp. 928–936, 1999.

[4] J. Bloit, N. H. Rasamimanana, and F. Bevilacqua, "Towards morphological sound description using segmental models," in *Proceedings of DAFx*, 2009.

[5] Axel Roebel, "A new approach to transient processing in the phase vocoder," in *6th International Conference on Digital Audio Effects (DAFx)*, London, United Kingdom, Septembre 2003, pp. 344–349.

[6] Axel Roebel and Xavier Rodet, "Real time signal transposition with envelope preservation in the phase vocoder," in *International Computer Music Conference*, Barcelona, Spain, Septembre 2005, pp. 672–675.

[7] Axel Roebel, "Shape-invariant speech transformation with the phase vocoder," in *InterSpeech*, Makuhari, Japan, Septembre 2010, pp. 2146–2149.

[8] B. Julesz, "Visual pattern discrimination," *Information Theory, IRE Transactions on*, vol. 8, pp. 84–92, 1962.

[9] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision*, vol. 40, no. 1, pp. 49–71, December 2000.

[10] J. H. McDermott and E. P. Simoncelli, "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis," *Neuron*, vol. 71, no. 5, pp. 926–940, Sep 2011.

[11] J H McDermott, M Schemitsch, and E P Simoncelli, "Summary statistics in auditory perception," *Nature Neuroscience*, vol. 16, no. 4, pp. 493–498, April 2013.

[12] Nicolas Saint-Arnaud and Kris Popat, "Computational auditory scene analysis," chapter Analysis and synthesis of sound textures, pp. 293–308. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 1998.

[13] Pierre Hanna and Myriam Desainte-catherine, "Time scale modification of noises using a spectral and statistical model," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003 IEEE International Conference on*, April 2003, vol. 6, pp. 181–184.

[14] Diemo Schwarz, *Data-Driven Concatenative Sound Synthesis*, Ph.D. thesis, Ircam - Centre Pompidou, Paris, France, January 2004.

[15] Joan Bruna and Stéphane Mallat, "Classification with invariant scattering representations," *CoRR*, vol. abs/1112.1120, 2011.

[16] Wei-Hsiang Liao, Axel Roebel, and Alvin W.Y. Su, "On stretching gaussian noises with the phase vocoder," in *Proc. of the 15th Int. Conference on Digital Audio Effects (DAFx-12)*, York, United Kingdom, September 2012.

[17] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[18] "http://anasynth.ircam.fr/home/english/media/experiment-texture-generation," .