A TWO LEVEL MONTAGE APPROACH TO SOUND TEXTURE SYNTHESIS WITH TREATMENT OF UNIQUE EVENTS

Seán O'Leary,
UMR STMS IRCAM - CNRS - UPMC
Paris, France
sean.oleary@ircam.fr

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

ABSTRACT

In this paper a novel algorithm for sound texture synthesis is presented. The goal of this algorithm is to produce new examples of a given sampled texture, the synthesized textures being of any desired duration. The algorithm is based on a montage approach to synthesis in that the synthesized texture is made up of pieces of the original sample concatenated together in a new sequence. This montage approach preserves both the high level evolution and low level detail of the original texture. Included in the algorithm is a measure of uniqueness, which can be used for the identification of regions in the original texture containing events that are atypical of the texture, and hence avoid their unnatural repetition at the synthesis stage.

1. INTRODUCTION

Sound textures are a class of sounds typically associated with the background of a scene that are somehow repetitive; for example rain, fire, or machinery. It is difficult to define precisely the properties of sound textures. Saint-Arnaud and Popat [1] offered some suggestions towards a working definition. They suggest that sound textures should, in some sense 'exhibit similar characteristics over time'; that is that one short snippet of a texture should exhibit similarities to another. They also suggest a two level description of textures. At the low level atoms of the texture are time localized sound elements, and the higher level describes the distribution of these atoms. They note that while such an atomic model is sometimes relevant to the physics of the texture, e.g. rain, they do not intend it as a general physical description. They give some points summarizing their working definition of sound textures.

1. Sound textures are formed of basic sound elements, or atoms.
2. Atoms occur according to a higher-level pattern, which can be periodic, random, or both.
3. The high-level characteristics must remain the same over long time periods (which implies that there can be no complex message).
4. The high-level pattern must be completely exposed within a few seconds attention span.
5. High-level randomness is also acceptable, as long as there are enough occurrences within the attention span to make a good example of the random properties.

McDermott et al [2, 3] suggest that given the temporal homogeneity of sound textures they can be characterized by time averaged statistics. This approach was inspired by previous work on image textures [4]. This hypothesis was tested by synthesizing various textures by imposing the statistics of a particular texture on a white noise sample. The statistics used described the amplitude envelopes of the textures after being passed through an auditory filterbank. These statistics included the first four moments of the envelopes, cross correlation between envelopes, and some measures relating to the autocorrelation of each envelope. The resulting synthesized sounds were not intended to be perceptually accurate reproductions, rather they were meant to test their hypothesis. They found that the synthesized sound textures could indeed be identified.

These studies give important insights into the requirements of a synthesis algorithm. There are many approaches to sound texture synthesis (for a thorough review of the literature see [5]). Broadly speaking, we can group these methods into model based approaches where the signal is synthesized from model parameters, and sampling or granular approaches where content from the original signal is used in the synthesized signal.

For many applications, such as cinema and computer games, realism of the synthesized sound is paramount. Sampling based methods can bring realism as they contain elements of the target sound. Some previous sampling based algorithms [6, 7] look for points of change to segment texture signals, these segments are then concatenated in a probabilistically determined sequence to produce the synthesized texture. The algorithm of Dubnov et al. [8] uses similarity in history and scale to select sampled wavelet coefficients. Drawbacks of sampling based methods include repetitions of parts of the original signal, difficulty modeling the higher level structure of the texture, and smooth concatenation of the sampled elements.

The proposed algorithm falls into the sampling based category. It looks to exploit regions of similarity in the original texture to inform the sequencing of sampled elements. There are two levels to the synthesis model. Longer term sections, called segments, are used to model the higher level structure of textures. These segments are synthesized from the concatenation of shorter term sections, called atoms. Atoms preserve the local structure of the texture. The sequences of both the segments and atoms are modeled probabilistically, this avoids repetition in the synthesized texture. A new overlap add method is introduced for concatenation. This enables concatenation with short overlap without introducing perceptible modulations.

The paper is organized as follows: Section 2 discusses the relationship of the algorithm to the properties of sound textures outlined in section 1. Section 3 presents the basic algorithm in detail while section 4 extends the algorithm to deal with unique events. Section 5 presents some sound examples. Section 6 presents some conclusions and possible future work.

2. THE RELATIONSHIP OF THE MONTAGE APPROACH TO SOUND TEXTURE PROPERTIES

As stated in the introduction, the montage approach to texture synthesis has two levels; segments and atoms. Segments are used to model the high level structure of the texture. By high level structure we mean features that determine the long term structure of a texture such as quasi-periodicity (e.g. pneumatic drill) or randomness (e.g fire). At the lower level atoms preserve the local structure of the segments.

Segments are modeled after longer sections of the texture. There is not a set length for a segment, rather they have user defined minimum and maximum lengths. The length of each segment is dependent on the selection of its successor. The sequencing of segments is informed by both local similarity for concatenation, and longer term similarity for preserving higher level structure. This sequencing has a probabilistic element to avoid repetition in the higher level structure of the synthesized texture.

These segments are used as templates for the synthesized texture. The segments are synthesized by a process of atom substitution. The original texture is split into atoms. These atoms all have the same user defined duration. For each atom a number of candidates are selected as possible replacements. These candidates are selected from throughout the texture based on the local similarity of the envelopes from an auditory filterbank analysis. This 'envelope matching' preserves the phase of envelope modulations in the synthesized texture. The synthesis of segments consists of substituting each of the original atoms with one of its qualifying candidates (including itself as one of the candidates). The selection of substitutes is probabilistic. This process preserves local structure and introduces new variation over the duration of the segment not present in the original texture. This is to avoid repetition on the atom scale in the synthesized texture.

The algorithm can be considered in terms of the properties of sound textures suggested by Saint-Arnaud and Popat [1] quoted in section 1.

- The presented model synthesizes textures from atoms.
- The high-level pattern of the atoms is preserved by sequencing them according to segments of the original texture. If there is periodicity in the texture it can be reproduced because the atoms will be aligned according to the original texture, this effectively matches the phase of the envelopes. Likewise randomness is maintained by randomizing both the selection of segments from the candidate successors and the choice of atom from the candidates for substitution.
- New high level structure will be introduced due to the sequencing of segments. As the long-term similarity of segments are matched this new structure should be coherent with the original texture.

The algorithm can also be considered in terms of the statistical description of the envelopes suggested by McDermott in [3].

- If the segments are distributed approximately evenly over the duration of the synthesis the moments of the envelopes will be approximately equal to those of the original.
- As the atoms are sampled from the original texture the local synchronicity of the envelope modulations is preserved. This is related to cross correlation of the envelopes in McDermott's texture model.
- The matching of atoms over localized time and frequency, the sequencing of atoms from segments of the original, and the transitions based on history all relate to the autocorrelation of the envelopes; the atom sequencing preserving local modulations and the segment sequencing preserving/synthesizing longer term modulations.

3. THE ALGORITHM

In this section the algorithm for analysis and synthesis of textures using the proposed approach is described. After the analysis phase the choices for synthesis are tabulated; all possible segments have candidates for their successors and each atom of the texture has candidates for substitution.

3.1. Analysis

The analysis stage of the montage approach involves finding regions in the texture that are in some way similar - this is necessary both for the selection of candidates for segment succession and the selection of candidates for atom substitution. The first step in the analysis is to represent the signal in a suitable form. As ultimately we are concerned with the perceptual closeness of the synthesized signal to the original a perceptually informed representation of the signal is utilized.

As much of the salient information in textures is contained in the envelopes of the auditory bands [3], a suitable comparison for similarity is taken to be a comparison of the time evolving energy from an auditory filter bank. The short time Fourier transform is a common and suitable processing platform, and so the algorithm will be presented in the context of the STFT.

The STFT is given by:

$$X\left(l,k\right)=\sum_{n=0}^{N-1}x\left(n\right)\omega\left(n-lh\right)e^{\frac{-i2\pi nk}{N}}. \tag{1}$$

Where l is the frame number, k is the frequency bin, N is the analysis window length and h is the hopsize.

Taking the envelopes to be the energies in subbands distributed according to the ERB scale:

$$engEnv_b\left(l\right)=\sum_{k=k_{b1}}^{k_{b2}}\left|X\left(l,k\right)\right|^2H_b\left(k\right). \tag{2}$$

Where k_{b1} is the first bin and k_{b2} is the last bin of the bth band and H is a bank of (frequency domain) band pass filters.

The envelopes then undergo further perceptual processing. The perceived change in loudness with intensity approximately obeys a power law. Hence the envelopes are compressed nonlinearly to simulate this. Each band is also scaled according to the equal loudness curve.

$$env_b=\left(engEnv_b/L\left(f_b\right)\right)^{0.3}. \tag{3}$$

Where L is the loudness curve, f_b is the centre frequency of the bth band and 0.3 is an experimentally determined exponent [9].

This gives a perceptually informed time/frequency representation of the signal sampled at the rate of the STFT analysis. Here we will refer to each time slice of both the STFT and the perceptually processed STFT as a frame.

The next stage in the analysis divides this representation of the signal into atoms. Each atom comprises several analysis frames

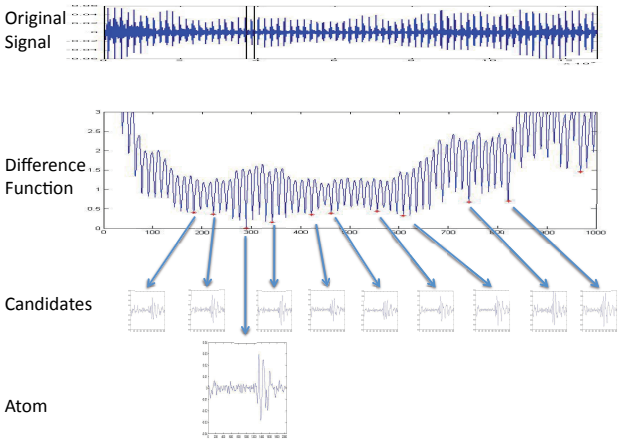


Figure 1: Selection of candidates for an atom

and have a 50% overlap with neighboring atoms. The atoms should be long enough to enable the comparison of envelopes and short enough to ensure enough variation in the synthesized texture. In our example set [10] we use $0.1s$ as the atom duration. This gives us a time/frequency representation of each atom. Each of these atoms undergoes further analysis; looking for similar regions over the duration of the texture.

3.1.1. Candidates for Atom Substitution

For each atom a difference function is created. This difference function gives us a measure of the difference between the atom under consideration and the associated region of the texture. The difference function for the ath atom at the lth frame is given by:

$$d_a(l) = \frac{\sqrt{\sum_{f=0}^{F-1}\sum_{b=1}^{B}\{env_b(l+f) - env_b(aF/2+f)\}^2}}{\sqrt{\sum_{f=0}^{F-1}\sum_{b=1}^{B}\{env_b(l+f)\}^2}}. \quad (4)$$

Where F is the number of frames in an atom and the atoms have a 50% overlap, i.e. an atom hopsize of $F/2$. This difference function is calculated at intervals of a single frame. This difference measure corresponds to the normalized euclidean distance between the auditory envelopes of the atom and the auditory envelopes of the texture from the lth to $l+F-1th$ frame.

A set of substitution candidates for each atom is selected from local minima in the difference function. There is a minimum time distance between selected candidates, dependent on the number of candidates to be selected. This is to ensure that candidates are selected from across the duration of the analyzed texture. This can be important for the selection of segments successors, as the candidates for substitution will also be considered as candidates for segment successors, and for this purpose it is desirable to have candidates spread over the duration of the texture.

An example of a difference function and candidates for a single atom of a texture are shown in Figure 1 for a helicopter sample (available to listen to at [10]). This is a quasi periodic texture, and this example illustrates how periodicity of events can be preserved with this model. Note that the envelope of the candidates is in phase with the envelope of the original atom. It is not necessary to retain the difference function after the analysis of an atom. Once the candidates for substitution are tabulated the difference function can be discarded. The result of the atom analysis is a list of pointers to the addresses in the original STFT of candidates for substitution and a normalized difference value for each of the candidates.

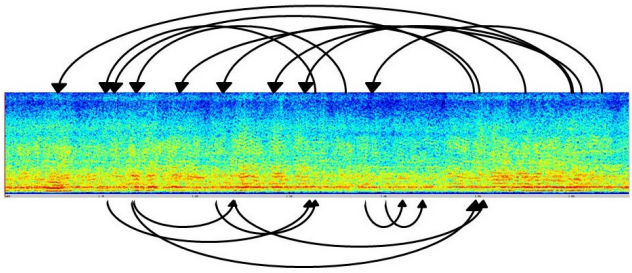


Figure 2: Transition from reading one segment to starting another

3.1.2. Candidates for Segment Successors

During synthesis segment succession occurs by substituting the last atom of the current segment with the beginning of its successor. And so each atom will be considered as a potential end of a segment and its candidates for substitution as a potential beginning for a succeeding segment. For segments, as well as the local similarity from the atom analysis, a longer term comparison is used. This is termed the history for the segment. Hence, a history comparison is also made between each atom and its candidates for substitution. This will be used to judge the possibility of a segment succession at the location of the atom during synthesis. No difference function is created as the history is only calculated for already found atom candidates.

As each segment has a minimum and maximum duration, the succeeding segment will begin between these points (see section 3.2.1). And so in the analysis phase each atom in this range is considered as a possible transition point from the current segment to its successor. This can be considered as a moving window analysis, the window length being the maximum minus the minimum duration of a segment. An example of a subset of possible segment succession points found for a texture is illustrated in Figure 2. For each step in this analysis there are typically many candidates. For example, for a single instance of this analysis if the difference between the minimum and maximum length between transitions is 1.5 seconds, and there are 20 atoms per second and 10 candidates per atom then there are 300 candidate points to consider as possible transition points. Only the succession points with the lowest measured difference are considered, again the selected succession points spanning the duration of the texture. The outcome of the succession analysis is a table of pointers for candidate segment end points for the current segment, associated difference values, and associated starting points for the next segment.

3.2. Synthesis

During synthesis the segment sequence is selected. From this the sequence of atoms is derived. These atoms are concatenated in the STFT domain before inverse Fourier transform and final overlap/add in the time domain are performed.

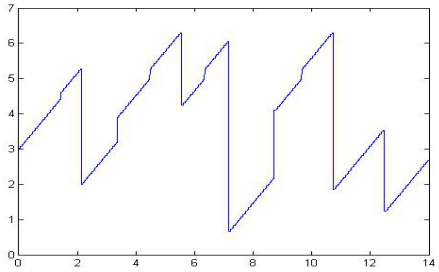


Figure 3: Sequence of transition in synthesized texture vs origin in original texture.

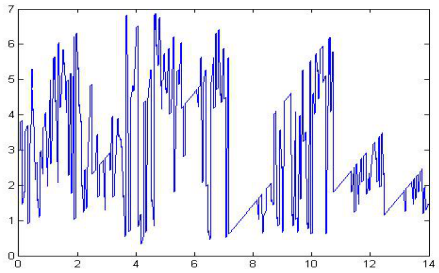


Figure 4: Sequence of atoms in synthesized texture vs origin in original texture.

3.2.1. Sequence Model

Starting from a random point in the original texture the algorithm selects successive segments from the candidates selected during analysis. This high level navigation of the texture acts as a template for the synthesized texture. There are some constraints on the choice of succeeding segments:

1. A segment must be at least a minimum, user defined, length.
2. A segment has a maximum, user defined, length.
3. If the succeeding segment occurs some time before the current segment in the original texture that time must be greater than a user defined minimum (at least equal to the length of the transition 'history').

The first constraint serves two functions: it prevents the synthesized texture from jumping too much and it allows the candidates for succeeding segments to be selected in the analysis phase. The second constraint prevents keeping the same high level structure as the original for long periods. The third prevents repeating parts of the high level structure in rapid succession.

Once a segment successor is selected the duration of the current segment is determined. The atoms for this segment can then be substituted probabilistically with the candidates selected during analysis, each of the qualifying candidates given equal probability of selection. A difference threshold can be used in the selection of atom substitutes. This defines the maximum difference allowed between atoms and possible switches. It was found that taking the median value of the normalized difference of all the candidates for all the atoms was an effective value for thresholding. An example of the sequencing of transitions and substitutions is illustrated in Figure 3 and 4.

The process of segment succession and atom substitution can continue for any desired period of time, producing varied textures which are perceptually similar to the original.

3.2.2. Overlap Add Operation

If we see the atoms as pieces of a jigsaw, the overlap-add operation can be seen to be a way of squeezing in pieces similar to the original into their place. Straightforward overlap-adding of broad band noise leads to modulations due to phase interference. Here a new solution to this problem is proposed. The cross fade of the atoms is done in the STFT domain. The number of frames involved in the cross fade is dependent on the bin number of the DFT (i.e. it is frequency dependent). The cross fade region is taken to be 4 times the inverse of the bin center frequency (i.e. 4 times the period), with a maximum of half the number of frames in an atom and a minimum of a single STFT frame. For bins with an overlap region less than half an atom length the point of maximum cross fade (i.e. 50%) is positioned at the point of least interference. This point is taken to be the point at which the absolute value of the complex difference in the overlap region is minimum.

4. DEALING WITH UNIQUE EVENTS IN THE TEXTURE

Often sampled textures contain local events that are uncharacteristic of the long term texture. Such events can be due to a recording artifact, an unwanted event in the recording, or a unique local event that is part of the process creating the texture. At the synthesis stage it may be desirable to avoid using atoms that contain such unique events as their repetition may be noticeable and artificial sounding in the synthesized texture; highlighting the sampling process and losing the naturalness of the synthesized texture.

Strobl [11] in a study of the concatenative algorithms of [6] and [7] refers to such events as 'disturbing elements', and proposes to identify them manually. Here we propose a method for identifying such elements that is a straightforward and natural extension to the montage approach.

There are two basic steps to this algorithm; 1) identify the unique region and 2) replace it with a qualifying piece of the texture. The replacement step allows the synthesis algorithm described above to remain unchanged.

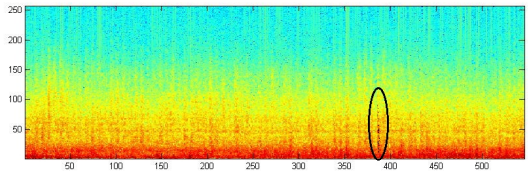
To identify events the difference measure obtained from 4 is utilized as a measure of the uniqueness of atoms. After the initial analysis stage each atom has a number of its closest matches from throughout the texture. The difference between an atom and its best match is taken to be a measure of its uniqueness.

A user defined parameter defines which atoms are to be replaced. This user parameter is a threshold and is stated as percentage of the maximum uniqueness found for the analyzed texture.

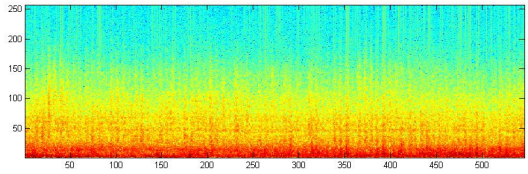
In order to find a region to replace the region selected as unique we again use the difference value defined by 4. Here we use a sum of the difference functions for the atoms adjacent to the selected region. The difference function of the latter atom in the sum is delayed by the appropriate time:

$$d_u(l) = d_{a1}(l) + d_{a2}(l + (w_u + 1)F/2). \quad (5)$$

Where $d_u(l)$ is the difference function used for finding the best match for the uth unique region, d_{a1} is the difference function for the $a1$th atom (the adjacent atom previous to the uth region) and d_{a2} is the difference function for the $a2$th atom (the adjacent atom following the uth region). The minimum of this function

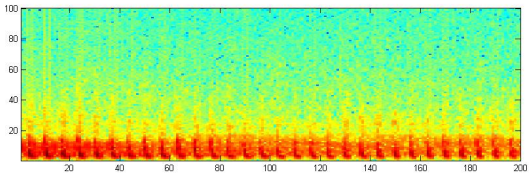


(a) Original with click highlighted.

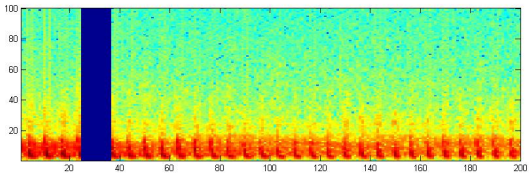


(b) Click identified and replaced.

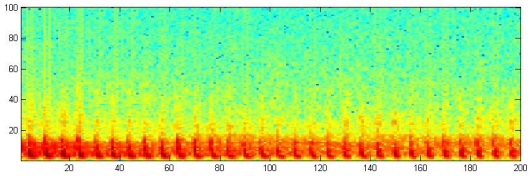
Figure 5: Example of identifying and replacing unique elements in a sampled texture.



(a) Original.



(b) With missing piece.



(c) Missing piece replaced.

Figure 6: Example of replacing a missing or damaged piece of a sampled texture (quasi-periodic helicopter).

gives the closest matching region, according to our measure, to replace the region identified as being unique. Once this region is identified the analysis described in section 3.1.1 is performed for the replacement atoms. Also, reference to the replaced atoms as substitutes for other atoms should be removed. Synthesis can then be performed as described in section 3.2.

An example of this is shown in figure 5. This is a recording of a steam train which contains a single 'click' sound. This 'click' is identified as the most unique region in the texture. For synthesis it is replaced as described above. This method can also be used to repair damaged recordings as is illustrated in figure 6. This illustration shows the spectrograms of helicopter sample, the same sample with a piece deleted, and the sample with the deleted piece replaced using the above method. Note how the approximate period of the events is preserved. The samples used to illustrate this are available at [10].

5. RESULTS

The presented algorithm was used to synthesize both textures containing quasi periodic elements and textures of a more random nature. The synthesized samples are twice the duration of the originals. The original samples were taken from [3]. The details of the synthesis for these sounds are as follows: the atom length was set to 0.1 seconds, the history set to 0.5 secs, and the maximum duration before a new transition set to 2 seconds. 20 candidates were selected for each atom, and 5 candidates selected for each transition. The transition candidates were selected by a simple sum of the normalized distance of the atom (local) difference and difference in histories. These examples can be found at [10].

6. CONCLUSIONS

In this paper an efficient and versatile algorithm for sound texture synthesis was presented. For efficient synthesis the atom and transition candidates can be tabulated from the analysis phase. Synthesis is then a fairly straightforward overlap add procedure in the STFT domain. The algorithm fulfills many requirements of a sound texture synthesis algorithm. At the low level the textures are synthesized from atoms and these atoms are sequenced to model the higher level organization of the original sound texture. Repetitions are avoided by introducing randomness in the sequencing of both the atoms and the segments, and smooth transitions are constructed by taking account of local similarity, longer history and a new overlap/add method.

While there are a number of user defined parameters in this algorithm, these parameters are not abstract, they have a natural relationship with the synthesis.

For the atom analysis the STFT hopsize determines the temporal resolution of the atom analysis, while the atom duration and difference threshold for substitution affect the variation of the timbre of the texture. For the segment sequencing the history length defines the region in which to compare the context of the high level structure, while the minimum and maximum length determine the high level variation.

The synthesis examples ([10]) show that for a large class of textures the synthesis is not extremely sensitive to these parameters. However, if there are extended events or a lot of variation in the original texture it may be beneficial to constrain the variation in the synthesized texture, i.e. lower the difference threshold for atom substitution and extend the history and minimum segment length.

As well as texture synthesis the algorithm has applications to editing textures, such as removing unique events or damaged portions of a sampled texture. The results seem very promising for a wide range of textures; from quasi periodic to random processes.

7. ACKNOWLEDGEMENTS

This research was funded by the French National Research Agency (ANR) as part of the PHYSIS project.

8. REFERENCES

[1] Nicolas Saint-arnaud and Kris Popat, "Analysis and synthesis of sound textures," in *in Readings in Computational Auditory Scene Analysis*, 1995, pp. 125–131.

[2] J H McDermott, A J Oxenham, and E P Simoncelli, "Sound texture synthesis via filter statistics," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-09)*, New Paltz, NY, Oct 18-21 2009, pp. 297–300, IEEE Signal Processing Society.

[3] 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.

[4] 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.

[5] Diemo Schwarz, "State of the art in sound texture synthesis," in *Intl. Conf. on Digital Audio Effects (DAFx-11)*, Paris, France, 2011.

[6] R. Hoskinson and D. Pai, "Manipulation and resynthesis with natural grains," in *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001.

[7] Lie Lu, Liu Wenyin, and Hong-Jiang Zhang, "Audio textures: theory and applications," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 2, pp. 156–167, 2004.

[8] Shlomo Dubnov, Ziv Bar-joseph, Ran El-Yaniv, Dani Lischinski, and Michael Werman, "Synthesis of sound textures by learning and resampling of wavelet trees," 2002.

[9] William M. Hartmann, *Signals, Sound, and Sensation*, Springer, New York, 1998.

[10] Seán O'Leary, "Sound examples," `http://anasynth.ircam.fr/home/english/media/montage-sound-texture-synthesis`.

[11] G. Strobl, "Parametric sound texture generator," M.S. thesis, Technische Universitääat Graz, Austria, 2007.