SOUND TEXTURE MODELING: A SURVEY

Gerda Strobl, Gerhard Eckel, Davide Rocchesso

SOUND TEXTURE MODELING: A SURVEY

Gerda Strobl, Gerhard Eckel
Institute of Electronic Music and Acoustics
University of Music and Dramatic Arts Graz
Inffeldgasse 10/3
8010 Graz, Austria
gerda.strobl@student.tugraz.at
eckel@iem.at

Davide Rocchesso
Department of Computer Science
University of Verona
Strada le Grazie 15
37134 Verona, Italy
rocchesso@sci.univr.it

ABSTRACT

Sound texture modeling is a widely used concept in computer music, that has its well analyzed counterpart in image processing. We report on the current state of different sound texture generation methods and try to outline common problems of the sound texture examples. Published results pursue different kinds of analysis /re-synthesis approaches that can be divided into methods that try to transfer existing techniques from computer graphics and methods that take advantage of well-known techniques found in common computer music systems. Furthermore we present the idea of a new texture generator framework, where different analysis and synthesis tools can be combined and tested with the goal of producing high quality sound examples.

1. INTRODUCTION

1.1. What is sound texture ?

Sound textures are an important class of sounds in interactive applications, video games, virtual reality and web-based applications, movie sound effects, or in extensive tracks of art installations. In video games it is important that sound textures can be used all over the game without requiring too much diskspace. In an installation-based scenario an unbound stream of audio creating a soundscape from a very short source material may be required.

Like in image processing [11] there is no universally valid definition of a sound texture. Since the term sound texture is relatively new several authors come up with *their* specific sound texture definition which is sometimes very vague and spans from baby crying and horse neighing up to background sounds with simple musical structure [1, 3, 6, 7, 8, 9, 13, 14, 18]. In the context of this paper we would like to adhere to the initial definition from Saint-Arnaud et al. who define sound texture using two major constraints: constant long-term characteristics and attention span.

A sound texture should exhibit similar characteristics over time. It can have local structure and randomness but the characteristics of the fine structure must remain constant on the large scale. A sound texture is characterized by its sustain[12, p 294].
This definition implies that pitch should not change dramatically (like with an accelerating car) and rhythm should neither accelerate nor slow down. *Attention span is the maximum time between events before they become distinct. High-level characteristics must be exposed within the attention span of a few seconds [12, p 298].*

From our point of view especially the second constraint is very interesting, regarding the difficulty of describing a sound having textural properties and the question of how many events have to happen in order to be able to differentiate between on single event and a continuous stream of events (e.g. a single car vs traffic).

When analyzing the examples of sound textures covered in most of the investigations on this subject, we can differentiate the following classes:
Natural sounds: fire, water (rain, waterfall, ocean) wind
Animal sounds: sea gulls, crickets, humming
Human utterances: babble, chatter
Machine sounds: buzz, whir, hammer, grumble, drone, traffic
Activity sounds: chip, sweep, rustle, typing, scroop, rasp, crumple, clap, rub, walking

In [7] also the sound of a crying baby is included. From our point of view this sound should not be regarded as a sound texture as the characteristics of the fine structure are not constant enough (c.f. definition quoted above).

Sound texture generation is at the intersection of many fields of research such as signal analysis, sound synthesis modelling, music information retrieval and computer graphics. Most of the published investigations pursue different kinds of analysis /re-synthesis approaches. Generally they can be divided into methods that try to transfer existing methods from computer graphics [3, 7, 9] and methods that take advantage of existing methods found in common computer music systems [14, 16] and speech synthesis [1, 9, 18]. Nevertheless all approaches start from a given audio sample and share the question of how to perform the best segmentation, which parameters to extract from the selected segments and how to do the best resynthesis in order to create a new sample longer in duration but with similar quality to the original. Apart from

[9] [1] , there are apparently no applications available today for producing sound textures, implying that - although the application area for sound textures seems very broad - the routinely production of sound textures nowadays is still based on manually editing recorded sound material.

Beyond basic sound demos Filatriau and Arfib try to develop a mapping between instrumental gestures and sonic textures. In [8] they describe their "textures scratcher", which is a digital music instrument employing a gesture - based exploration of a visual space. It consists of a Max/MSP adaption of the Functional Iteration Synthesis algorithm presented in [16].

2. STATE OF THE ART

2.1. Analysis and Synthesis of Sound Textures (1998)

In [14] sound texture is understood as a two-level phenomenon, having a low-level (atoms) and a high-level basis (distribution and arrangement of the atoms). A cluster-based probability model (k-means) that encodes the most likely transitions of feature vectors, is used to characterize the high-level of sound textures. The resynthesis is done by a binary tree structured quadrature mirror filter bank (QMF).

2.2. Synthesis of Environmental Sound Textures by Iterated Nonlinear Functions (1999)

Di Scipio [16] uses nonlinear functions to synthesize environmental sound textures. The presented system is not based on any signal analysis but on nonlinear function iteration using the method Functional Iteration Synthesis (FIS). A sine wave is used as the iterated discrete form. Sounds like rain, cracking of rocks, burning materials etc. are considered as sonic phenomena of textural nature. Sounds that are obtained from sine iteration feature an internal, chaotic fluctuation which are "slower" than white noise. They show amplitude and phase modulation curves that result into a micro-level activity.

2.3. Manipulation and Resynthesis with Natural Grains (2001)

A method is presented for extracting parts of an audio signal in order to reproduce it in a stream of indeterminate length [9]. The created sounds are not referred to as sound texture but the used natural sounds correspond with the sound examples from other investigations in sound texture. Also, the analysis/synthesis approach is derived from findings in image texture synthesis. A wavelet transform [2] is computed on windowed frames of the input signal in order to fragment the material into syllable-like segments. Then an automatic speech recognition algorithm determines natural transition points. The sound in between the transition points is not broken up any further. These segments are called *natural grains*. For each segment a table of similarity between it and all the other segments is constructed. After the segmentation a first-order Markov chain is used where each segment is corresponding to a state of the chain. The transition probabilities from one state to the other are estimated based on the smoothness of transition between it and all the other segments. Then the segments are arranged in a continuous stream with the next segment being chosen from the other segments which best follow from it.

2.4. Synthesizing Sound through Wavelet Tree Learning (2002)

A statistical learning algorithm is presented for synthesizing sounds that are statistically similar to the original [7]. An input sound file is decomposed into wavelets. Out of the wavelets the algorithm captures the joint statistics of the coefficients across time and scale. Then a multiple resolutions analysis tree is created that is used to create new collections of sound grains that have a similarity to the original sample sound. In the re-synthesis step the inverse wavelet-transform is applied to obtain an output tree. By random granular combination the texture sounds are re-synthesized. This approach is used for "periodic" (ocean waves) and stochastic (crying baby) sound textures. This approach is directly derived from prior work of the authors in texture movie synthesis.

2.5. Sound texture modelling with linear prediction in both time and frequency domain (2003)

Sound texture is considered apart from music and speech signals as a third class of sounds [1]. In this work texture is modelled as rapidly modulated noise by using two linear predictors in cascade. The first linear prediction operation is applied in the time domain in order to capture the spectral envelope. The second linear prediction is carried out in the frequency domain which uses the residual of the previous LPC analysis to estimate the temporal envelope. In the resynthesis step a filtered Gaussian noise is used to feed the cascade of filters whose coefficients were obtained by the analysis of the original texture sample. Finally frames are overlapped to create a continuous signal.

2.6. Sound Texture Modelling and Time-Frequency LPC (2004)

Similar to the approach presented in [1], this paper [18] applies the time frequency LPC method to create a generative sound model. The major goal is to synthesize arbitrarily long audio streams that are perceptually similar to the original sound. After the frequency domain (FDLPC) computation the event density over the entire frame is calculated as a statistical feature of the sound texture and is used in the synthesis process to control the occurrence of events. In a further step the detected events are extracted, leaving a background sound devoid of any events. The

[1] `http://www.cs.ubc.ca/~reynald/applet/Scramble.html`

[2] The wavelet filter can be chosen in the corresponding Java applet.

individual segments are concatenated and a time domain TDLPC filter is applied to the background sound to model it. The obtained TDLPC coefficients are used to reconstruct the background sound in the resynthesis process. In a next step the time and the frequency domain LPC coefficients are clustered to reduce amount of analysis data. In the resynthesis process the background sound and the event sequence are generated separately and mixed subsequently. For a required re-synthesis duration, a noise excited background filter is used to generate the background sound. To generate the foreground sound the event density number is used as the parameter of a Poisson distribution to determine the onset position event in the re-synthesized sound. Finally the events are re-synthesized using a random subclip index.

2.7. Creating Audio Textures by Samples: Tiling and Stitching (2004)

This approach attempts to make longer sounding sound textures from short input samples [3]. Starting from image processing, existing methods for creating visual textures (tiling and stitching) are transferred to the sound domain. The tiling-based method uses a chaos mosaic to generate a new sequence from a sound texture sample whereas the stitching-based method combines multiple chunks using a least recently used algorithm.

3. SOUND EXAMPLES

The most important listening observations are:

- All the available generated texture samples that are available online can be recognized as a re-synthesized version of the original sample and show strong similarity.
- The type of sample can be recognized without knowing the original.
- The synthesized samples contain audible repetitions: evident repetition of blocks (e.g. crowd [3][3]) and implausible accentuations that create an undesired rhythmical pattern (e.g. waves) [7][4]).
- The important events are very well synthesized but the background sounds appear blurred (e.g. fire[1][5]).
- In some examples gaps of silence can be heard, that make the samples sound unnatural and disturb the notion of homogeneity typical for the original recording (e.g. traffic [3]).

[3] http://pages.cpsc.ucalgary.ca/~parker/AUDIO/

[4] http://www.cs.huji.ac.il/labs/cglab/papers/texsyn/sound/

[5] http://www.ee.columbia.edu/~marios/ctflp/ctflp.html

4. RELATED WORK

The approaches that are presented in the following subsections show similarities to the investigations listed in the chronology but differ because of different sound examples that are used (e.g. rhythmical music [10]), additional effects that are added (e.g. mixture of modeled sound and real recording[4, 5]) and a completely different background (e.g. music texture as a technical term in musicology) that nevertheless shares the ”texture” term.

4.1. Audio Textures

Audio Texture [12] is introduced as a new medium, as a means of synthesizing long audio streams from a short example audio clip. The example clip is segmented using a window function. Then Mel frequency cepstral coefficients are computed in order to get a feature vector for every segment. The similarity and the transition probability between any two frames are computed one to the other. From the similarity measurement a novelty score is computed to extract the audio structure and to segment the original audio in a new order. According to their definition audio textures includes sounds like lullabies, simple background music and natural sounds (e.g. horse neighing, sea shore).

4.2. Music Textures

From a short example clip of music an infinite version is generated by changing the order of the frames of the original [10]. The new medium *music texture* is inspired from *video texture* [15] and is understood as a collection of sounds. The approach is essentially based on a metrical representation (downbeat, meter, etc.) and on grouping by similarity.

We should mention that the term music texture is also used in reference to the overall structure of a piece of music, the number of parts playing at once, the timbre of the instruments playing these parts as well as the harmony and the rhythmic structure used in a piece of music. The formal terms that are used, describe the relationships of melodies and harmony (e.g. monophony, polyphony, etc.).

4.3. Sound Textures based on Computational Fluid Dynamics

In [4, 5] methods are developed for real-time rendering of aerodynamic sounds and turbulent phenomena (swinging swords and fire). Vortex sounds are simulated corresponding to the motions of fluids obtained by the simulation. The process can be considered as a virtual recording process of the sound. First numerical fluid analysis is used to create the sound textures. Then the texture samples are used for rendering the vortex sound corresponding to the motion of the fluids. As these sounds do not only consist of vortices (e.g. combustion, reverb), a combination of vortex and recorded sounds is used.

4.4. Granular Synthesis

Granular synthesis presents probably one of the oldest approaches in computer music to create texture like sounds. On purpose we say "texture like sounds" because with grain based methods not every type of sound texture can be (re)-synthesized but generally we would like to mention that all the presented methods that start from an input sample use a special form of granular synthesis in their final synthesis step. A recent approach focusing on the generation of natural noisy sounds is the GMEM Microsound Universe (GMU) [2] which consists of a collection of Max/MSP objects that allow the generation of high density streams (e.g. sound of rice falling on a plate, rain drops) with a precise parameter control.

5. FUTURE WORK

We are planning to build a parametric sound texture generator that allows for creating sounds out of a simple parameter control structure. Starting from a short input sequence, different, new, and unconstrained texture sounds of variable length will be produced. These new sequences should consist of consecutively connected patterns that are similar to the input sequence.

The generator will be based on an analysis / re-synthesis approach and will be implemented with the graphical language pure data (PD) [6] allowing for different approaches of analysis and re-synthesis to be tested. First we will test existing algorithms with regard to the resulting sound quality. In a further step we will create a collection of analysis/synthesis PD-abstractions that can be combined and have control inputs. A reason for using pure data is its real time computation ability which is important to allow for an interactive modification of the analysis parameters and methods. The goal of the implementation is an application for demonstration purposes (see Figure 1) allowing the user to manually adjust parameters until the produced sound texture sounds plausible and realistic with respect to a particular use of the texture. It will also be possible to add transformation effects in the generator-interface (e.g. slight pitch-shifting, time-stretching, change of envelope and amplitude etc.) in order to improve the overall sound quality. The real-time implementation will include the possibility to adapt (e.g. change of window length) the analysis process to the source material used such that the re-synthesized signal approximates the original signal as closely as possible. Eventually also the chosen re-synthesis method will depend on the input signal. Although all input examples are considered as sound textures they usually show very different properties.

6. CONCLUSIONS

Although the category of sound textures is difficult to define with precision, we think that by taking the definition from [14] and by reviewing various examples from the literature, we can contribute to a better comprehension of sound textures. We have presented a review of current investigations on sound texture generation and show that - although sound texture synthesis produces similar results to the input samples - the sound quality cannot be considered convincing. Finally we propose a future model of a real-time sound generator, that consists of an analysis/synthesis framework including the possibility to test and compare different approaches. The main goal of our approach is to avoid audible repetitions and to improve the perceptual quality of the generated sound textures.

As a special issue we would like to focus on the improvement of the quality of the sound results in order to create sound textures that can be used in practical applications. We specially want to emphasize that repetitions should not be audible and sound textures should be targeted of sounding perceptually "meaningful", in the sense that the synthesized texture is perceptually comparable to the example clip. In the ideal case, no difference should be noticeable, i.e. the generated sounds still sound natural and contain no artefacts. Furthermore we would like to find out whether it is possible to create good results with a single method incorporating control structures or if we could improve the result by implementing a model-based generator (e.g. modeling of crackling sounds by using a special Poisson distribution [17]) and modify both the analysis and synthesis step for to input signal characteristics.

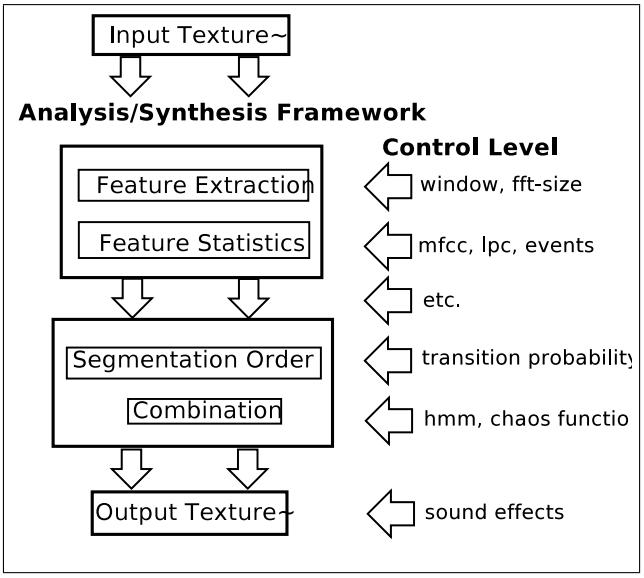

Figure 1. Sketch of a possible framework, having a fixed internal analysis/synthesis structure that can be controlled and adapted from the control functions.

7. ACKNOWLEDGEMENTS

We wish to thank Antonio de Sena, Pietro Polotti, Federico Fontana and Georg Holzmann for their very helpful comments and for proof-reading this paper.

[6] http://www.puredata.org

8. REFERENCES

[1] M. Athineos and D. Ellis. "Sound texture modelling with linear prediction in both time and frequency domains", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '03*, vol. 5, pp. 648-51, 6-10 April 2003.

[2] C. Bascou and L. Pottier. "GMU, A Flexible Granular Synthesis Environment in Max/MSP", *Proceedings of the Sound and Music Computing Conference 2005*, Salerno, 2005.

[3] B. Behm and J.R. Parker. "Creating Audio Textures by Samples: Tiling and Stretching", *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP '04*, vol.4, pp 317-320, 17-21 May, 2004.

[4] Y. Dobashi, T. Yamamoto and T. Nishita. "Real-time Rendering of Aerodynamic Sound using Sound Textures based on Computational Fluid Dynamics", *ACM Transaction on Graphics*, vol. 23(3),pp. 732-740, 2003.

[5] Y. Dobashi, T.Yamamoto and T. Nishita. "Synthesizing Sound from Turbulent Field using Sound Textures for Interactive Fluid Simulation" *Proceedings of Eurographics* , vol. 23(3),pp. 539-546, 2004.

[6] S. Dubnov and N.Tishby. "Analysis of Sound Textures in Musical and Machine Sounds by means of Higher Order Statistical Features", *Proceedings of the International Conference on Acoustics Speech and Signal Processing*, Munich, 1997.

[7] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski and M. Werman. "Synthesizing Sound through Wavelet Tree Learning" *IEEE Computer Graphics and Applications*, vol. 22, no. 4, pp. 3848, Jul/Aug 2002.

[8] J-J. Filatriau and D. Arfib. "Instrumental Gestures and Sonic Textures", *Proceedings of the Sound and Music Computing Conference 2005*, Salerno, 2005.

[9] R. Hoskinson and D. Pai. "Manipulation and Resynthesis with Natural Grains", *Proceedings of the International Computer Music Conference ICMC01*,Havana, Cuba, 2001.

[10] T. Jehan. *Creating Music by Listening*, PhD thesis, Massachusetts Inst. Technology, Cambridge, 2005.

[11] F. Liu. *Modeling Spatial and Temporal Texture*, PhD thesis, Massachusetts Inst. Technology, Cambridge, 1997.

[12] L. Lu, L. Wenyin and H. Zhang., "Audio Textures: Theory and Applications", *IEEE Transactions on Speech and Audio Processing*,vol. 12, pp 156-167, March 2004.

[13] M. J. Norris and S. L. Denham. *Sound Texture Detection using Self-Organizing Maps*, Center for Theoretical and Computation Neuroscience, University of Plymouth, UK, November 2003.

[14] N. Saint-Arnaud and K. Popat. "Analysis and Synthesis of Sound Texture", *Computational Auditory Scene Analysis*, D. F. Rosenthal, Horoshi G. Okuno,editors, Lawrence Erlbaum Association, New Jersey 1998.

[15] A. Schödl, R. Szeliski, D. H. Salesin and I. Essa. "Video textures", K. Akeley, editor, *Siggraph 2000, Computer Graphics Proceedings*, pp 3342. ACM Press, ACM SIGGRAPH, Addison Wesley Longman, 2000.

[16] A. Di Scipio. "Synthesis of Environmental Sound Textures by Iterated Nonlinear Functions", *Proceedings of the 2nd COST g-6 Workshop on Digital Audio Effects DAFX'99*, Trondheim, Norway, 1999.

[17] Sethna and Paul A. Houle. "Acoustic Emission from crumpling paper" *Physics Review E*, vol. 54(1),pp 278-283, 1996.

[18] X. Zhu and L. Wyse. "Sound Texture Modelling and Time-Frequency LPC" *Proceedings of the 7th Int. Conference on Digital Audio Effects DAFX'04*, Naples, 2004.