Sound texture modeling
Research Proposal for the
PhD in Sound and Music Computing

Stefan Kersten

March 2009

1 Introduction

Many sounds that surround us and that we perceive in our daily lives have textural properties – yet sound texture it is a term difficult to define, because the sounds subsumed are usually not perceived consciously and a definition might also strongly depend on the context the sound is perceived in.

Sound textures do convey information, and yet they exhibit some of the statistical properties that are normally attributed to noise. This section is an attempt to introduce the notions of sound texture within its broader context, the soundscape, sum up definitions previously expressed in the literature and provide a "working definition" of the notion of sound texture that shall be the subject of this research.

1.1 Soundscape and sound texture

In his book "The Soundscape" [33], R. Murray Schafer describes the notion of "acoustic ecology" as our level of awareness of our acoustic environment [40]. He groups the sounds that make up our sonic environment into three broad categories:

Keynote sounds are background sounds that define our acoustic and emotional environment, similar to keynotes in music defining the fundamental underlying tonality. Keynote sounds are not listened to consciously, but perceived unconsciously.

Sound signals are foreground sounds that are intended to attract attention and generally bear a meaning to be transported to the listener.

Soundmarks are sounds that have –in analogy to landmarks– a particular meaning for a community and its visitors.

Were we to undertake a preliminous classification of "sound texture", it would be found in the category of *keynote sounds* rather than *sound signals*, while *soundmark* intuitively describes a broader concept that might contain sounds from both categories.

One of the earliest attempts of defining the term "sound texture" can be found in N. Saint-Arnaud's Master thesis "Classification of Sound Textures" [31] and later in a book chapter in "Computational Auditory Scene Analysis" [32]. Saint-Arnaud assumes a sound texture model to be comprised of two elements:

Sound atoms are basic building blocks in an arbitrary (often time-frequency) feature space that make up the textural sound source.

Transition rules between atoms that model the statistics of atom (co-)occurrence in feature space.

Furthermore, an important characterization of sound texture implies its evolvement over time: Saint-Arnaud and other authors require the signal to exhibit the properties of a stationary statistical process on a certain time scale [41, 2]. Textures are thus characterized by their perceptual and statistical properties and not by individual events [31]. Saint-Arnaud explicitly limits the time window in which the signal

is to be regarded stationary to the "attention span", i.e. the amount of time that can pass before two events are perceived to be distinct, which lies in the proximity of a few seconds.

Zhu [41] defines sound textures as "[...] sounds for which there exists a window length such that the statistics of the features measured within the window are stable with different window positions", which again refers to a stationary signal, and seems to be the most intuitive and practical definition.

In extension of this notion, the class of sounds subject of this research shall be described as multi-variate, multi-resolution time series, where the signal might be a superposition of *various independent* stationary stochastical processes, each exhibiting stationarity on a potentially different time scale. Being able to specify or predict the extent to which the model to be developed is able to capture the statistics of this signal mixture is one of the goals to be pursued.

In order to get an impression of the sounds to be modeled, Table 1 lists some examples.

Applause
Running water
Waves
Rainfall
Fire
Birdsong
People babble
Machine sounds
Traffic noise and urban ambiances

Table 1: Examples of sound textures.

2 Motivation

Schafer's categorization suggests, that sound texture transports information not in an information theoretic sense, but rather as a carrier of emotional and situational percepts. Indeed, sound texture –often denoted atmosphere– forms an important part of the sound scene in real life, in movies, games and virtual environments in general.

The ability to model texture in a statistical sense, without detailed knowledge or assumptions about the source material, leads to several desirable properties that a texture model should possess:

Compactness of representation The model should require (significantly) less parameters than the original coded audio.

Statistical properties The signal statistics should be discoverable using a limited amount of training data.

Research in sound texture modeling has been promising, but it also leaves open some questions and poses new problems:

- Sound texture research is still lacking a formal framework for subjective evaluation of the perceptual qualities, and in fact the minority of previous approaches have been rigorously evaluated, neither subjectively nor objectively.
- Recent findings in multi-layer, perceptually relevant signal models have not been systematically incorporated in sound texture modeling, a deficiency that is all the more substantial, as in general a particular source signal model cannot be assumed, due to the statistical nature of the sound material.
- Recent research contributions to highly expressive generative models for multi-variate time series has not yet found its way into statistical modeling of sound signals.
- The extension of sound texture modeling to capture the full perceptual fidelity of a binaural or multichannel sound field is only in its beginnings.

- The possibility of capturing various source textures in a single model and provide a meaningful, data-driven parameterization of the synthesis has not been systematically explored yet.

In the following, the goals and possible contributions of this proposed research are laid out, followed by a list of prospective applications and the concrete embedding of the research in a funded project of applied research.

2.1 Research goals and contributions

This proposal is intended to address a number of problems still open in this area of research.

The first priority is to develop a systematic framework for perceptual evaluation of sound texture models in the context of virtual environment sonification. This is an important step in order to compare existing and future sound texture models in terms of their performance during synthesis.

Another important contribution is to extend the previously used signal models by more sophisticated, perceptually motivated signal decompositions in the hope of modeling the signal statistics in a feature space closer to human perception. Closely related is the development of a more capable statistical model in feature space –based on recent advances in time series models, see 3.3– with the intention of being able to model larger-scale interdepencies of features and thus extending the types of sounds that can be modeled in a perceptually meaningful way.

The application of the probabilistic model to the task of sound texture classification is a logical step, given that a generative model can be optimized for classification with relatively minor modifications [14].

2.2 Applications

Generative sound texture models are applicable in diverse fields of research and applications.

The generation of perceptually convincing soundscapes for Virtual Reality applications is an important means of conveying a sense of presence and immersion for the user. Parameterized models, that are capable of synthesizing soundscapes according to the situational and geographical position of an avatar, can potentially provide a sense of movement and presence in the virtual environment, that is hard to achieve with methods based on audio sample playback. In this vein, sound texture models can help in ambience generation in computer games and sound installations, where often the textural background needs to be provided for an indefinite amount of time while data storage capacities are limited.

Textural properties are not only present in sound, but also in images and more generally in many physical processes. An appropriate parametric texture model could be a building block in sonification and auditory display applications.

In electroacoustic music and soundscape composition, sound textures and their parametric manipulation play an important role [38]. Generative, probabilistic sound texture models can not only be creatively employed by composers in their work, but can also aid in the automated analysis and classification of such music.

In the related field of Computational Auditory Scene analysis, statistical models can be a useful tool for sound scene description and classification and database indexing. Similarly, classifiers derived from a generative model, can be employed as a tool in model-based scene analysis [10, 11].

Statistical models provide an extreme form of perceptual data reduction and can be employed in abstract level audio coding applications [2] as well as in audio restoration tasks [21].

The researched proposed here is to be conducted within the bigger framework of the three-year research project *Metaverse1*[1], that has the general goal of developing standards for emerging virtual environments. The research at the Music Technology Group is focused on automatic soundscape generation from a community-built corpus of synthetic and environmental sounds, that are tagged with user-supplied textual keywords[2]. Sound texture modeling plays an important role, not only in transporting a certain sustained sense of "auditory scene realism", but also in reducing the amount of data needed for each scene to be modeled. Sound texture models certainly cannot account for the entire interactive soundscape, simply because the amount of deterministic interaction is naturally limited, but they can provide a background for otherwise more prominent "signal" sounds found in the acoustic virtual environment.

[1] `http://www.metaverse1.org/`
[2] `http://www.freesound.org/`

3 State of the art

The relatively sparse amount of works related to sound texture modeling is contrasted by a vast amount of research in image and video texture synthesis, which also had a direct influence on some of the methods proposed for sound texture modeling.

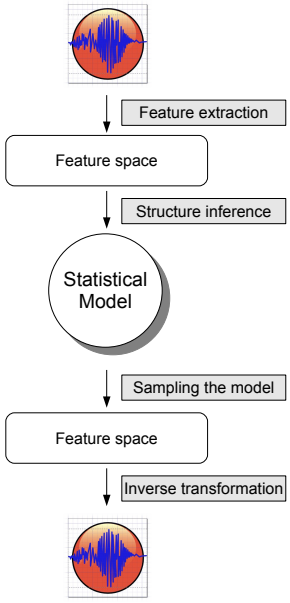

Figure 1: Architecture of a typical sound texture modeling system.

A depiction of a general sound texture modeling framework is shown in Figure 1. A preprocessing and feature extraction step is followed by a structural inference algorithm, that learns a model of the signal statistics in feature space. Generating new, random textures that are statistically –and ideally also perceptually– similar to the source texture involves *sampling the model* to generate random vectors in feature space according to the belief the model has of the signal statistics and transforming those feature vectors back to the time domain to obtain a PCM coded audio sample.

Research in computational modeling and synthesis of sound textures has started in the late 1990's and the signal analysis –or feature extraction– approaches can be categorized in three different categories:

Methods based on multiresolution signal decomposition The multiresolution wavelet transform is used to extract signal features at multiple time and frequency resolutions.

Methods based on time-domain signal decomposition Atomic grains are extracted from the source signal and recomposed in the time domain.

Methods based on a source-filter signal model Source filter analysis –more specifically linear predictive coding (LPC)– is used to analyse the signal in both the time and the frequency domain.

In section 3.1, previous work on sound texture modeling is reviewed.

The statistical models employed so far in sound texture modeling have been comparatively inexpressive; often the signal statistics are modeled with simple Markov chain models on the feature vectors. Recently, progress has been made in training deep hierarchical generative models that can adequately capture the fast varying statistics in multidimensional time series, e.g. raw video frames. Section 3.3 reviews some of these models and relates them to the ones used in current methods of sound texture modeling.

Other relevant research includes works on presence in virtual environments and sound texture description and classification in general, some of which are summarized in section 3.4.

3.1 Sound and image texture synthesis

3.1.1 Methods based on multiresolution signal decomposition

Multi-resolution analysis (MRA), in particular the discrete wavelet transform (DWT, see Figure 2), is potentially well suited to modeling the dynamics of sound textures, where important perceptual detail is present in various frequency bands and on different time scales. Using quadrature mirror (QMF) lowpass and highpass filters, the signal is recursively split in low-resolution approximation coefficients and high-resolution detail coefficients, which are obtained by applying the wavelet transfer function $\Psi(z)$.

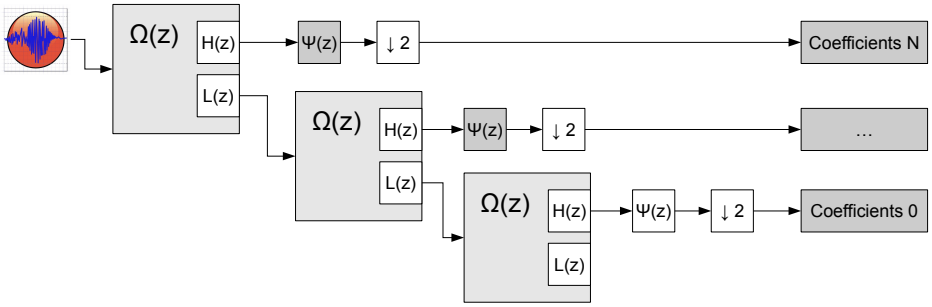

Figure 2: Filterbank interpretation of the multi-resolution discrete wavelet transform.

One of the earliest attempts of modeling sound textures for synthesis is described in [32]. Although the authors do not use the wavelet transform, they employ a simple filter bank of octave-spaced quadrature mirror filters to split the signal into frequency bands. Feature vectors are extracted by grouping amplitude features in the frequency bands in time and thus capturing the temporal evolvement of the signal by transitions of features. The feature vectors are then clustered by a probabilistic algorithm (K-means) to model the high-dimensional probability mass function (PMF). During synthesis, the one-dimensional PMF conditioned on the preceding feature values is estimated from the clusters and an appropriate feature vector is selected and transformed to the time domain to form the resulting texture.

Bar-Joseph et al. [3] analyse static image and time-varying video textures by means of a hierarchical, multi-resolution analysis, where wavelet coefficients of different temporal and spectral scales are represented as an N-ary tree. Based on work by Basseville [4], the statistical model assumes that the tree was generated by a stationary statistical process and models paths in the tree to a given hierarchy by a simple linear sequence model. A new tree is generated by sampling from the path model and successively adding layers, starting from the low-resolution signal components and gradually filling in the high-resolution details. Bar-Joseph at al. also describe the synthesis of textures with a mixture of features from two source textures by statistically merging the analysis trees of the texture sources and generating from the merged tree.

Dubnov et al. [7] apply wavelet analysis to sound texture synthesis. They extend the idea of learning conditional probabilities along paths in a wavelet tree by also learning the predecessor probabilities of nodes in a given hierarchy level, thus estimating the statistics of the temporal order of coefficients. The algorithm introduces a parameter ϵ that provides a threshold for the distance computation when determining path prefixes for the Markov chain statistics and which determines the "randomness" of the signal, or, conversely, its similarity to the input signal.

In [24], Miner and Caudell [3] describe a wavelet synthesis technique for realistic texture –stochastic, non-pitched– sounds. The DTW is used to analyse the signal on different time and frequency scales, and various textural sounds are obtained from a single source analysis by parametric modification of wavelet coefficients before resynthesis by means of the inverse wavelet transform (IWT). Interestingly, Miner followed an iterative design process, in which modifications in the synthesis model are followed by systematic perceptual evaluation by means of listening experiments (see 3.4); however, no attempt was made to model the statistics of the wavelet coefficients.

[3] This work was preceeded by Miner's PhD thesis [23], which unfortunately wasn't available for this assessment.

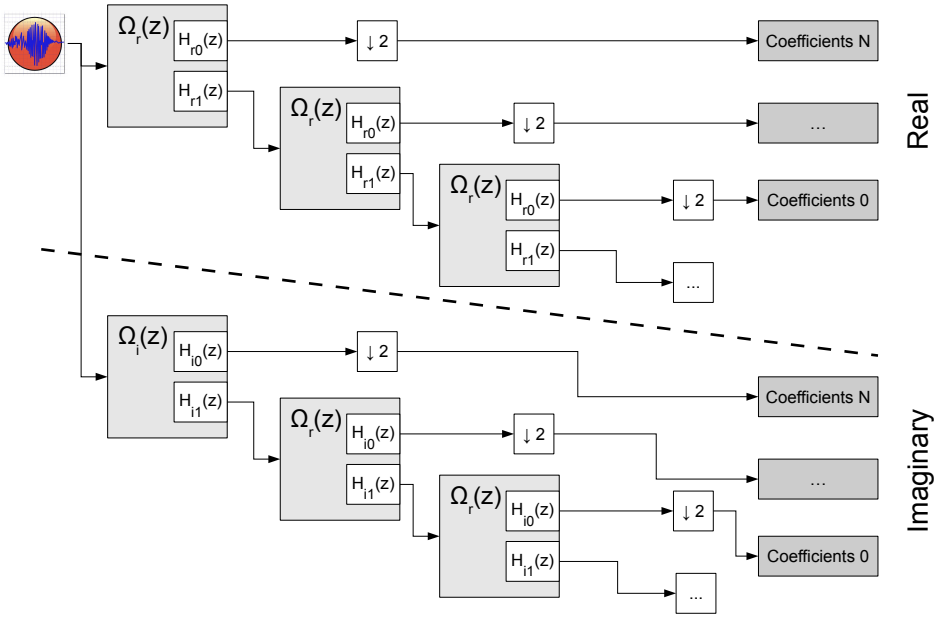

Figure 3: Filterbank interpretation of the dual-tree complex wavelet transform.

O'Regan and Kokaram [27] propose the use of the Dual-Tree Complex Wavelet Transform (DTCWT, see Figure 3), which has the desirable property of shift-invariance [19]. They model the sound texture statistics by a Markov-Random-Field (MRF) on the wavelet tree coefficients, similar to the approach taken by Efros and Leung for image texture synthesis [9]. The authors report improved results compared to the ones obtained by [7], although selecting parameters for the synthesis algorithm seems crucial for good results.

In [26], Misra et al. describe an analysis and resynthesis system for entire soundscapes. A source sound is decomposed into its transient and sinusoidal components [34] and the residual is modeled by wavelet tree learning as described in [7]. Transient events and harmonic components are not modeled stochastically, but can be manipulated in a user interface for later recomposition and resynthesis.

3.1.2 Methods based on time-domain signal decomposition

In this section we subsum sound texture models for synthesis, that attempt to decompose the signal on a single time scale, segmenting accross the whole frequency band and recomposing the resultant pieces according to some statistical model.

In [16] and later in [17] Hoskinson describes a sound texture synthesis model, that first segments the audio signal into "natural grains" –loosely corresponding to phonemes in speech– based a speech segmentation algorithm proposed by Alani and Deriche [1]. From the extracted grains a first-order markov chain of grain transition probabilities is built according to a simple estimate of "smoothness of transition" between pairs of grains. The synthesis process then samples from the markov chain and concatenates grains into the output stream.

In [21], Lu et al. develop a method of *audio texture synthesis* based on spectral and temporal features extracted from the audio signal. The idea of *audio texture* is inspired by video textures, and refers to sounds that are "relatively monotonic in nature" and which can be used to generate infinite streams of audio for game and background music applications, and for filling transmission gaps in audio restoration. The proposed method first extracts Mel-Frequency Cepstral Coefficients (MFCC) from the audio signal and attempts to infer patterns –or "subclips"– from a frame similarity matrix based on an algorithm borrowed from music information retrieval research [12]. MFCC vectors for new audio textures are sampled from the frame and subclip similarity matrices and are –after an optional step of time- and frequency-scale modification– transformed into the time domain and concatenated.

Other time-domain decomposition and resynthesis approaches include [28], where equal length "chunks" extracted from the audio signal are concatenated according to a least-squares similarity measure and preventing repeated reuse of the same chunk by means of a least-recently used (LRU) algorithm. [18] maps human motion gestures obtained by video feed analysis to a "sonic grain space", that enhances an achitectural space by a sound texture in an installative setting. Hoffman et al. [15] describe a recomposition procedure for musical audio, where a feature extraction stage (MFCC) is followed by a statistical analysis based on Dirichlet-Process Hidden Markov Models (DP-HMMs) –allowing to model a large sparse state space– and an extension that allows to incorporate the features of multiple musical sequences into a single model.

3.1.3 Methods based on a source-filter signal model

A very different approach to sound texture modeling involves the assumption of a source filter signal model.

Athineos and Ellis [2] consider sound textures to be

- *Noisy*, i.e. without strong periodic components
- *Rough*, i.e. amplitude modulated in the 20-200Hz range

They model a source sound by first estimating the spectral power envelope by Linear Predictive Coding (LPC) analysis [22], whitening the signal by inverse filtering with the obtained IIR filters and estimating the temporal envelope of audio signal by performing an LPC analysis in the frequency domain –a procedure called Temporal Noise Shaping (TNS) [13]. This implicit extraction of transients and microtransients –captured in the temporal envelope– is claimed to be essential for modeling the fine temporal microrstructure present in many sound textures: "The technique has greatest success with sounds that include both broadband noise and densely-packed microtransients. Such sounds are very difficult to represent by methods that detect and separate transients from the rest of the residual." In order to assess the quality of the resynthesis, the authors introduce an error metric based on the short time Fourier transform (STFT) magnitude.

Zhu and Wyse [41] also use the Time-Frequency LPC (TFLPC) to model sound textures. In contrast to Athineos and Ellis' approach, they estimate event on- and offsets from a salience function based on the LPC filter gain derivative. Events are extracted from the source signal and the "background din" is concatenated and modeled separately. A further data reduction step for modeling events is performed by converting the polynomial time- and frequency-domain LPC coefficients to the reflection domain and clustering by use of K-means. During resynthesis, event onsets are modeled by a Poisson distribution and event feature vector sequences are drawn from the distributions estimated by the cluster means and variances. Event sequence are transformed separately and mixed in the time domain to obtain an output texture.

3.2 Signal models

Recent advances in phase vocoder signal analysis (Figure 4) improve the detection and classification of spectral peaks in the phase vocoder frequency spectrum. In [29], a new method for transient detection and processing in the phase vocoder is described.

The transient detection scheme operates in the frequency domain and is based on the determination of the Center of Graphity (COG) of a spectral peak within the analysis window [6]. A probabilistic model, sampled from the signal, is used to adaptively derive the threshold for the assignment of peaks to transients based on the COG.

In [30], the same author derives spectral peak feature descriptors from a STFT analysis:

- *Frequency coherence* as the peak distance from the bin center
- *Energy location* within the anlysis window
- *Time duration*
- *Normalized bandwith*

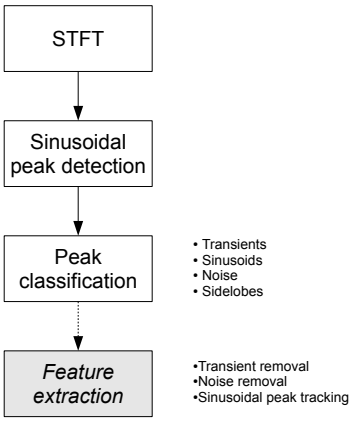

Figure 4: Phase vocoder block diagram.

The extracted features are then used in a decision tree classifier, that has been built by hand from the analysis prototypical signal (Figure 5).

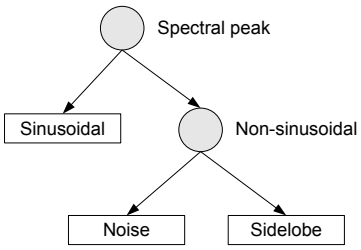

Figure 5: Peak classification in the phase vocoder.

3.3 Generative models for multivariate time series

Many time series data –in particular audio data– contain detail on multiple temporal and spectral scales. This section reviews some statistical approaches to modeling multi-variate, multi-resolution time series.

Basseville et al. [4] provide a theoretical formulation of a statistical, multi-resolution signal model, by assuming a generative statistical source, that, when sampled, generates a signal statistically similar to the source signal. They show that stationary signals can be represented and generated in a hierarchical fashion, such that the low-resolution parts of the signal are generated first, and the finer details are generated with a probability, that only depends on the low-resolution details in the proximity of the current hierarchy level. This hierarchical representation was successfully applied to modeling the statistics of two-dimensional image textures.

One possible strategy to model multiscale input is to design hierarchical feature extractors, where the modules on different hierarchy levels model features on different scales and levels of abstraction. Bengio and LeCun [5] provide an interesting discussion from a statistical point of view why hierarchical architectures are indispensable when dealing with complex, multi-variate data.

Recently, there has been a surge of research in hierarchical distributed representations, due to the discovery by Hinton et al. that certain layered architectures can be learned efficiently in a greedy manner, one layer at a time [14].

Taylor et al, [37] use a modification of Hinton's approach to model time-series data, in this case skeletal angles of human motion tracking recordings, that are highly non-linear in the observed feature space of joint angles.

The basic building block of the multi-layer architecture is the Restricted Boltzmann Machine (RBM, see Figure 6), a two-layer neural network with logistic units and non-directed between-layer connections (but no within-layer connections). The RBM comprises an undirected energy-based probabilistic model

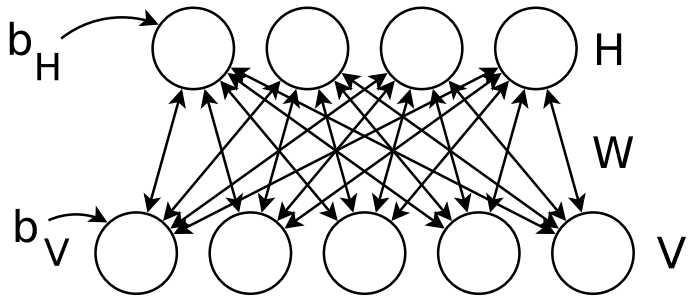

Figure 6: Restricted Boltzmann Machine (RBM).

(as opposed to a directed belief-net), where the probability of a hidden logistic unit h_j being "on" given the visible data vector v is:

$$p(h_j = 1|v) = f(b_j + \sum_i v_i w_{ij})$$

where f is the logistic function, b is a bias term and w is a symmetric connection weight matrix. Taylor et al. model visible data vectors by a Gaussian distribution of variance 1, where the mean is estimated by the bias term c:

$$p(v_i|h) = \mathcal{N}(c_i + \sum_j h_j w_{ij}, 1)$$

Maximum likelihood learning is slow in RBMs, but [14] showed that following the gradient of another function called contrastive divergence works well in practice:

$$\Delta w_{ij} \propto \langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}$$

where $\langle \cdot \rangle$ is the expected value with respect to the data distribution and the distribution of the reconstruction, respectively.

Taylor et al. extend the single RBM model to a Conditional Restricted Boltzmann Machine (CRBM) that is able to model temporal dependencies. The model adds directed connections from visible variables in the previous n time steps to the visible state (bias) in the current time step and similarly undirected connections from previous visibles to the current hidden state. Given the data at time $t, t-1, ..., t-n$ the hidden units at time t are conditionally independent and the CRBM can be trained by contrastive divergence.

The autoregressive weights from previous visibles to current visibles can model short-term temporal structure well, while the hidden units capture longer-term, higher-level structural information. This effect can be improved by stacking modules of CRBMs similar to Deep Belief Networks [14], where hidden units are treated as fully observed variables in higher layers.

Taylor et al. report their model to be able to capture the features of various different motion sequences, where transitions either come from recorded transitions in the training data, or can be approximated by adding noise to the hidden variables during the reconstruction in order to encourage the model to explore different regions of the feature space.

A similar multi-layer stacking of temporal RBMs is described in [35], where it is used to model the statistics of raw video pixel frames. The model is able to capture the highly non-linear dynamics of a physical simulation of bouncing balls; due to the deep architecture a significantly lower number of parameters has to be optimized compared to a Hidden Markov Model.

3.4 Evaluation methodologies

In this section some previous research on determination of the perceptual properties of sound texture and the perceptual validation of synthesis results shall be reviewed.

Saint-Arnaud [31] provides informal textual descriptions of sound textures and a list of some possible perceptual properties, determined in a series of "brainstorming" sessions in group meetings.

Miner and Caudell [25] evaluate the results of wavelet based sound texture synthesis with a number of listening experiments. The listening test design encompasses similarity rating, freeform identification and context-based rating experiments. The data obtained from the experiments was analyzed using data mining techniques like multidimensional scaling (MDS) and Pathfinder analysis, and the results were used in turn to modify the synthesis algorithm in a perceptually informed manner.

Witmer and Singer [39] present questionnaires for assessing the sense of presence in virtual environments. They define "presence" in terms of the underlying concepts "involvement" and "immersion":

- "Involvement is a psychological state experienced as a consequence of focusing one's energy and attention on a coherent set of stimuli or meaningfully related activities and events."
- "Immersion is a psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences."
- "Presence is defined as the subjective experience of being in one place or environment, even when one is physically situated in another. [...] Both involvement and immersion are necessary for experiencing presence."

Based on these definitions, they develop two questionnaires: The Presence Questionnaire (PQ) measures the degree of presence of an individual in a virtual environment and the influence of the contributing factors. The prerequisite Immersive Tendencies Questionnair measures the tendency of an individual to be involved or immersed. Both questionnaires employ a seven-point scale format based on semantic differential principle [8].

4 Framework and methodology

The goal of this research is to develop a perceptually sound, adaptive statistical model of a certain class of sounds called *sound textures* that exhibit properties of a stationary process on a certain timescale. In contrast to CASA systems, that try to model the statistics one or several sound sources in the presence of noise, the aim is to model the joint statistics of all sound sources present in the mixture.

Previous research in sound texture modeling has focused on signal models, that could only explain one aspect of the signal properties, such as wavelet tree analysis or LPC methods for mostly stochastic sounds or time-domain methods focusing on clearly separable atomic sound events. This proposed research tries to answer two basic questions (from which a number of others arise):

- Can the use of a layered signal model help in reducing the number of correlations in feature space to be modeled by the statistical model?
- Can a more powerful statistical model significantly improve the quality of synthesis and extend the range of sounds that can be modeled?

Recent advances in the representational power of generative models [14, 35, 37, 36] suggest that those models can be successfully applied to the statistics of features in a psychoacoustic signal model. The proposed architecture is shown in Figure 7. One branch of evaluation should be performed comparatively, i.e. comparing synthesis results with real textures; the goal is to develop a listening experiment framework that focuses on the perceptually relevant features of the sound textures being modeled. Another important evaluation criteria is the performance of the developed models in the context they are being employed in; thus, another branch of evaluation should be the performance assessment in the context of a virtual environment soundscape, based on the presence questionnaires developed by [39].

To summarize, the cornerstones of the proposed research are:

- Multi-layer signal model
- Multi-layer generative model for signal statistics

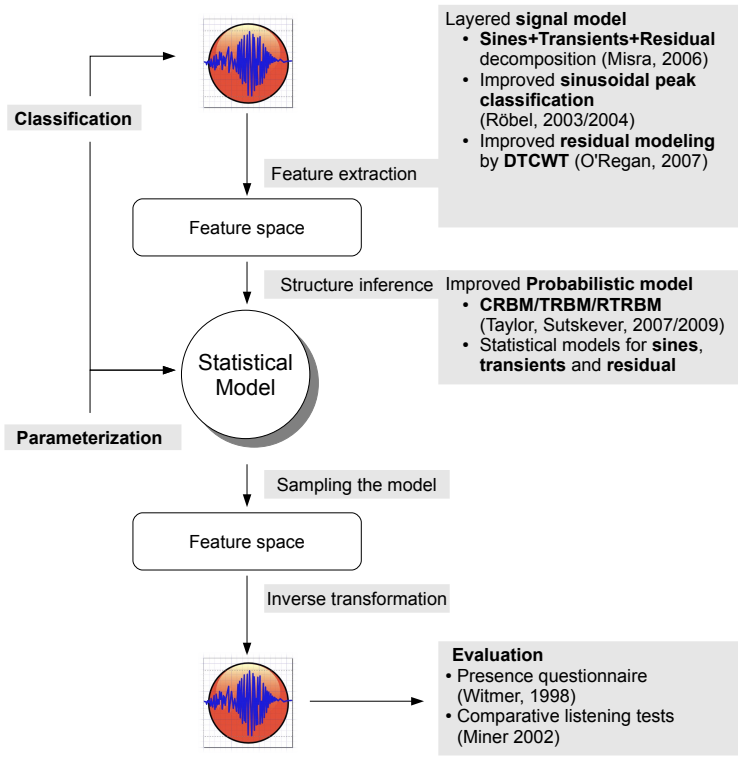

Figure 7: Proposed sound texture model.

- Systematic evaluation of synthesis results
- Application to multi-texture modeling
- Application to texture classification

In the following an approximate research plan, divided in annual research goals is laid out:

- First year
  - Develop an evaluation framework
  - Implement and evaluate the most promising existing approaches ([26] and [27]) within this framework
  - Extend the signal model in [26] with the improved residual modeling from [27]
  - Improved signal decomposition by sinusoidal peak classification [30]
  - Evaluate the resulting models
- Second year
  - Model signal statistics with a deep architecture (CRBM, TRBM, RTRBM)
  - Extend the model to multiple texture modeling and parameterization
  - Evaluate the resulting models
- Third year
  - Apply the generative model to classification tasks and evaluate the performance
  - Final evaluation of the resulting models

References

[1] A. Alani and M. Deriche. A novel approach to speech segmentation using the wavelet transform. In *Signal Processing and Its Applications, 1999. ISSPA '99. Proceedings of the Fifth International Symposium on*, volume 1, pages 127–130 vol.1, 1999.

[2] M. Athineos and D.P.W. Ellis. Sound texture modelling with linear prediction in both time and frequency domains. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, volume 5, pages V–648–51 vol.5, 2003.

[3] Ziv Bar-Joseph, Ran El-Yaniv, Dani Lischinski, and Michael Werman. Texture mixing and texture movie synthesis using statistical learning. *IEEE Transactions on Visualization and Computer Graphics*, 7(2):120–135, 2001.

[4] M. Basseville, A. Benveniste, K.C. Chou, S.A. Golden, R. Nikoukhah, and A.S. Willsky. Modeling and estimation of multiresolution stochastic processes. *Information Theory, IEEE Transactions on*, 38(2):766–784, 1992.

[5] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Machines*, volume to appear. MIT Press, 2007.

[6] Leon Cohen. *Time-frequency analysis: theory and applications*. Prentice-Hall, Inc., 1995.

[7] S. Dubnov, Z. Bar-Joseph, R. El-Yaniv, D. Lischinski, and M. Werman. Synthesizing sound textures through wavelet tree learning. *Computer Graphics and Applications, IEEE*, 22(4):38–48, July 2002.

[8] Robert Dyer, J. J. Matthews, J. F. Stulac, C. E. Wright, and K. Yudowitch. Questionnaire construction manual. annex: Literature survey and bibliography. Technical report, Operations Research Associates, July 1976.

[9] Alexei A. Efros and Thomas K. Leung. Texture synthesis by Non-Parametric sampling. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, page 1033. IEEE Computer Society, 1999.

[10] Daniel P. W. Ellis. *Prediction-driven computational auditory scene analysis*. PhD thesis, Massachusetts Institute of Technology, June 1996.

[11] Daniel P. W Ellis. Model-based scene analysis. In D. Wang and G. Brown, editors, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, page 115–146. Wiley/IEEE Press, 2006.

[12] J. Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (ICME 2000)*, volume 1, pages 452–455, 2000.

[13] J. Herre and J. D. Johnston. Enhancing the performance of perceptual audio coders by using temporal noise shaping (TNS). In *Proc. 101st AES Convention*, 1996.

[14] G. E Hinton, S. Osindero, and Y. W Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

[15] M. D. Hoffman, P. R. Cook, and D. M. Blei. Data-driven recomposition using the hierarchical dirichlet process hidden markov model. In *Proc. 2008 International Computer Music Conference*, 2008.

[16] R. Hoskinson. *Manipulation and Resynthesis of Environmental Sounds with Natural Wavelet Grains*. PhD thesis, The University of British Columbia, 2002.

[17] Reynald Hoskinson and Dinesh K Pai. Synthetic soundscapes with natural grains. *Presence: Teleoperators & Virtual Environments*, 16(1):84–99, February 2007.

[18] Joanne Jakovich and Kirsty Beilharz. ParticleTecture: interactive granular soundspaces for architectural design. In *Proceedings of the 7th international conference on New interfaces for musical expression*, pages 185–190, New York, New York, 2007. ACM.

[19] Nick Kingsbury. Complex wavelets for shift invariant analysis and filtering of signals. *Applied and Computational Harmonic Analysis*, 10(3):234–253, May 2001.

[20] M. S. Lewicki. Efficient coding of natural sounds. *Nature Neuroscience*, 5(4):356–363, 2002.

[21] Lie Lu, Liu Wenyin, and Hong-Jiang Zhang. Audio textures: theory and applications. *Speech and Audio Processing, IEEE Transactions on*, 12(2):156–167, 2004.

[22] J. D. Markel and A. H. Gray. *Linear Prediction of Speech*. Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[23] Nadine E. Miner. *Creating wavelet-based models for real-time synthesis of perceptually convincing environmental sounds*. PhD thesis, The University of New Mexico, 1998.

[24] Nadine E. Miner and Thomas P. Caudell. A wavelet synthesis technique for creating realistic virtual environment sounds. *Presence: Teleoper. Virtual Environ.*, 11(5):493–507, 2002.

[25] Nadine E. Miner, Timothy E. Goldsmith, and Thomas P. Caudell. Perceptual validation experiments for evaluating the quality of wavelet-synthesized sounds. *Presence: Teleoper. Virtual Environ.*, 11(5):508–524, 2002.

[26] Ananya Misra, Perry R. Cook, and Ge Wang. A new paradigm for sound design. In *Proc. of the 9th Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 319–324, Montreal, Quebec, Canada, September 2006. http://www.dafx.ca/proceedings/papers/p_319.pdf.

[27] D. O'Regan and A. Kokaram. Multi-Resolution sound texture synthesis using the Dual-Tree complex wavelet transform. In *Proc. 2007 European Signal Processing Conference (EUSIPCO)*, 2007.

[28] J.R. Parker and B. Behm. Creating audio textures by example: tiling and stitching. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, volume 4, pages iv–317–iv–320, 2004.

[29] Axel Röbel. Transient detection and preservation in the phase vocoder. In *Proc. 2003 International Computer Music Conference*, 2003.

[30] Axel Röbel, Miroslav Zivanovic, and Xavier Rodet. Signal decomposition by means of classification of spectral peaks. In *Proc. 2004 International Computer Music Conference*, pages 446–449, Miami, USA, 2004.

[31] Nicolas Saint-Arnaud. *Classification of Sound Textures*. Master thesis, Massachusetts Institute of Technology, 1995.

[32] Nicolas Saint-Arnaud and Kris Popat. *Analysis and synthesis of sound textures*, pages 293–308. L. Erlbaum Associates Inc., 1998.

[33] R. Murray Schafer. *The Soundscape: Our Sonic Environment and the Tuning of the World*. Destiny Books, 1994.

[34] Xavier Serra. Musical sound modeling with sinusoids plus noise. In Curtis Roads, Aldo Piccialli, Giovanni De Poli, and Stephen T. Pope, editors, *Musical Signal Processing*, page 480. Swets & Zeitlinger, 1997.

[35] I. Sutskever and G. Hinton. Learning multilevel distributed representations for High-Dimensional sequences. In *Proc. Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, San Juan, Puerto Rico, March 2007.

[36] I. Sutskever, G. Hinton, and G. Taylor. The recurrent temporal restricted boltzmann machine. In *Proc. 21st Conference on Neural Information Processing Systems (NIPS 2008)*, Vancouver, B.C., Canada, 2009.

[37] Graham W. Taylor, Geoffrey E. Hinton, and Sam T. Roweis. Modeling human motion using binary latent variables. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, page 1345–1352. MIT Press, Cambridge, MA, 2007.

[38] Barry Truax. Soundscape composition as global music: Electroacoustic music as soundscape. *Organised Sound*, 13(02):103–109, 2008.

[39] Bob G. Witmer and Michael J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators & Virtual Environments*, 7(3):225–240, June 1998.

[40] K. Wrightson. An introduction to acoustic ecology. *Soundscape: The Journal of Acoustic Ecology*, 1(1):10–13, 2000.

[41] X. Zhu and L. Wyse. Sound texture modeling and timefrequency LPC. In *Proc. of the 7th Int. Conf. on Digital Audio Effects (DAFX-04)*, Naples, Italy, October 2004.