

Smooth Granular Sound Texture Synthesis by Control of Timbral Similarity

Diemo Schwarz, Sean O’Leary

Ircam–CNRS–UPMC

firstname.secondname@ircam.fr

ABSTRACT

Granular methods to synthesise environmental sound textures (e.g. rain, wind, fire, traffic, crowds) preserve the richness and nuances of actual recordings, but need a preselection of timbrally stable source excerpts to avoid unnaturally-sounding jumps in sound character. To overcome this limitation, we add a description of the timbral content of each sound grain to choose successive grains from similar regions of the timbre space. We define two different timbre similarity measures, one based on perceptual sound descriptors, and one based on MFCCs. A listening test compared these two distances to an unconstrained random grain choice as baseline and showed that the descriptor-based distance was rated as most natural, the MFCC based distance generally as less natural, and the random selection always worst.

1. INTRODUCTION

The synthesis of credible environmental sound textures such as wind, rain, fire, crowds, traffic noise, is a crucial component for many applications in computer games, installations, audiovisual production, cinema. Often, sound textures are part of the soundscape of a long scene, and in interactive applications such as games and installations, the length of the scene is not determined in advance. Therefore, it is advantageous to be able to play a given texture for an arbitrary amount of time, but simple looping would introduce repetition that is easy to pick out. Using very long loops, or layering several loops can avoid this problem (and is the way sound designers currently do this), but this stipulates that a long enough recording of a stable environmental texture is available, and uses up a lot of media and memory space.

We present here a method to extend an environmental sound texture recording for an arbitrary amount of time, without the need for the source recording to be of a stable and uniform timbre or density. This means, a sound designer can use a recording that fits the scene in atmosphere, but without needing to isolate a stable and sufficiently long loop, since our method will ensure smooth timbral transitions, while still varying the texture to avoid repetition effects.

Our method is based on randomised granular playback with control of the similarity between grains using two different timbral distance measure that are compared in an evaluation: a timbral distance based on audio descriptors, and an MFCC-based distance. We also compare to purely randomised playback as a baseline.

2. PREVIOUS AND RELATED WORK

The method presented here situates itself in the granular synthesis-based approaches to sound textures, as opposed to ones based on signal or physical models. These methods need a recording as source material from which sound grains are picked and played back. Granular playback takes advantage of the richness of actual recorded sound, in contrast to other methods based on pure synthesis [4], see the state-of-the-art overview on sound texture synthesis [19] for further discussion and a general introduction of sound textures. Fröjd and Horner [5] use purely randomised playback of long grains (around one second), with half-grain crossfade, and slight randomisation of playback parameters (detuning, amplification) to avoid repetition. O’Leary and Roebel’s *Montage approach* [14, 15] exchanges grains by timbral similarity to avoid repetition, while following template sequences from the original, and introduce a spectral crossfade minimising phase distortion.

Specifically, the present research draws on previous work on corpus-based sound texture synthesis [20, 21], that can also be seen as content-aware granular synthesis, and extends the work of Fröjd and Horner [5] by the explicit modeling and control of timbral similarity on randomised granular playback. Other methods to extend a given texture are based on modeling of higher-order statistical properties [1, 9, 11, 12]. All these latter methods need a source recording with stable and uniform texture content while our proposed method can work with more varied textures by being aware of the timbral content of all grains.

Other methods for sound textures go further by modeling and recreating the typical transitions occurring in the source texture by wavelet- or Markov-trees [3, 7, 8].

A recent approach by Heittola et al. [6] quite similar to ours is aimed at full soundscape synthesis to recreate the acoustic environment of a specific location for digital maps. There, the timbral similarity is calculated on MFCCs and their deltas averaged over four second grains. The resulting similarity matrix serves to coalesce adjacent grains into longer segments, and to cluster these in order to control the smoothness of transitions.

3. TEXTURE SYNTHESIS

Our method is derived from corpus-based concatenative synthesis (CBCS) [18], where grains are played back from a corpus of segmented and descriptor-annotated sounds. Usually, CBCS is used to control the timbral evolution of the synthesised sound while still using original recordings as the sound source. This can be applied to texture synthesis to match the sound to the evolution of a given scene [20], see also the example video of interactive wind texture synthesis in a 2D descriptor space¹, when the descriptor target is given directly by the sound designer, or by the game engine. However, in the application described here, we don't want to control the timbral output directly, but have the system synthesise a varying texture without audible repetitions nor artefacts such as abrupt timbral or loudness changes. To this end, we use a timbral distance measure d between the last played grain and all other grains as candidates, and randomly select a successor grain from the timbrally closest grains, thus generating a random walk through the timbral space of the recording, that never takes too far a step, but that potentially still traverses the whole space of expression of the recording.

The algorithm proceeds as follows:

1. We construct a corpus of one or more recordings, segment it into grains (here of length 800 ms without overlap), and analyse each grain i for its timbral characteristics in a feature vector u_i .

In our experiments we used two variants of annotation giving rise to two different distance measures:

- (a) An analysis of the 7 audio descriptors validated by [20], extracted with the IRCAMDESCRIPTOR library [16]: The mean of the instantaneous descriptors *Loudness*, *FundamentalFrequency*, *Noisiness*, *SpectralCentroid*, *SpectralSpread*, *SpectralSlope* over all frames of size 23 ms.
 - (b) An analysis of the timbral shape in terms of the mean of the mel-frequency cepstral coefficients (MFCCs) over the segment.
2. For synthesis, we start with a seed grain q , selected randomly or given manually to start off with a certain timbral content.
 3. When a grain is triggered, $c = 5$ successor grains are searched by a $(c + 1)$ -nearest neighbour search, i.e., given the current grain's descriptor values u_q as query point, the k D-tree [2, 22] finds in logarithmic time the c candidate grains with descriptor values closest to the query (and the query grain q itself, since it has a distance of zero). The distance function is a weighted Euclidean distance, with weights given by the inverse standard deviation to normalise the search space. Multiplying the weights allows us further to give more importance to certain descriptors, or to exclude them from influencing the search.

4. The successor grain s is chosen randomly from the c candidate grains. If s is within one second of q , a new grain s is picked from the candidates, to avoid picking grains too close to each other.
5. To avoid too regular triggering of new grains, the duration and time of the next grain are randomly drawn within a 600–1000 ms range, and a random start offset of ± 200 ms is applied to each grain.
6. Played grains are overlapped by 200 ms, and an equal-power sinusoidal cross-fade is applied during the overlap.
7. While the desired length of the output texture is not reached, the chosen grain s becomes the query grain q , and the algorithm continues at step 3.

3.1 Implementation

The prototype system is implemented in Max/MSP using the MuBu (Multi-Buffer) extension library [17], with the integration of the batch analysis module pipo.ircamdescriptor².

4. RESULTS AND EVALUATION

The method presented here is evaluated in an ongoing listening test accessible online³. At the time of writing, 31 subjects took the test.

The test consists of a questionnaire with 7 second extracts of 7 sound examples listed in table 1. This small test database contains sounds from [3] that are widely used in evaluation of sound textures [5, 10] and thus partially allows comparison of the results. Other sounds were contributed by [20] and by the partners of the PHYSIS project⁴. All sounds were chosen for their properties of being a non-uniform environmental sound texture, i.e. containing some variation in texture and timbre, but not clearly distinguishable short sound events. An exception is the Baby Crying sound, that is here as an extreme counterexample, since it contains very different and well-separated cries.

Sound Example	Description
Lapping Waves	long-term structure
Desert Wind	wind with occasional gusts
Stadium Crowd	atmosphere, occasional cheering and honking
Water Faucet	various speeds of water flow
Formula One	not actually a texture, containing structured variation
Traffic Jam	motor sounds, honking, some shouts
Baby Crying	not actually a texture, containing large variation

Table 1. List of Test Sound Database and description

¹ <http://imtr.ircam.fr/imtr/Sound.Texture.Synthesis>

² <http://forumnet.ircam.fr/product/max-sound-box>

³ <http://ismm.ircam.fr/sound-texture-transition-control-evaluation>

⁴ <http://sites.google.com/site/physisproject>

	orig	descr	mfcc	random
Lapping Waves	85.09 (\pm 20.70)	73.04 (\pm 19.76)	71.82 (\pm 23.58)	46.01 (\pm 25.16)
Desert Wind	92.46 (\pm 08.70)	59.90 (\pm 22.28)	61.97 (\pm 26.19)	23.24 (\pm 23.11)
Stadium Crowd	91.65 (\pm 13.36)	56.22 (\pm 29.05)	23.03 (\pm 18.52)	25.83 (\pm 20.18)
Water Faucet	86.82 (\pm 16.93)	55.38 (\pm 24.01)	25.34 (\pm 18.11)	14.18 (\pm 15.11)
Formula One	95.15 (\pm 08.57)	29.55 (\pm 20.62)	17.43 (\pm 19.61)	12.85 (\pm 15.71)
Traffic Jam	77.36 (\pm 26.44)	59.01 (\pm 31.47)	56.23 (\pm 29.07)	52.97 (\pm 27.07)
Baby	95.43 (\pm 09.45)	17.98 (\pm 15.15)	13.07 (\pm 15.38)	15.89 (\pm 21.02)
Total	89.14 (\pm 17.30)	50.16 (\pm 29.76)	38.41 (\pm 31.37)	27.28 (\pm 26.15)

Table 3. Naturalness rating mean and standard deviation over all subjects.

For each example, the original, and 4 test stimuli of 7 s length are presented. The stimuli contain in randomised order the 3 syntheses (by descriptor distance, MFCC distance, random), and the original as hidden reference. For each stimulus, the subject is asked to rate the aspect of *Naturalness* on a scale of 0–100, with labels given in table 2. Note that the question of *Sound Quality* does not make sense for this evaluation since no signal processing other than long cross-fades is applied, and therefore the perceived sound quality is the same for all stimuli.

We linearly scaled the collected naturalness ratings individually for each subject (over all sounds) to a range of 0 to 100. The rationale is that the relative ratings of overly enthusiastic or overly critical subjects are thus made comparable with the rest of the subjects. We can see in figures 3 and 4 that only a few subjects (notably 1, 12, 14, 27) exhibit very narrow rating ranges.

The collected data is summarised in figure 1. We can see that the descriptor-based similarity measure generally obtains better ratings than the MFCC based one, that the random grain choice is rated worst, and that the originals are rated very high, with only a few outliers.

To test if the observed differences of means are significant or simply due to chance, further statistical analysis has been carried out using the ANOVA (analysis of variance) method with Bonferroni correction. Here the null hypothesis H_0 is that means are equal, and differences are due to chance, and the alternative hypothesis H_A is that the means are not equal. The p-values and significance levels⁵ for each pair of comparisons are given in tables 4 and 5 for the raw and scaled ratings, respectively. The scaling seems to augment the contrast of the results, leading to a rise in significance level for a few pairs in the ANOVA results.

ANOVA confirms that globally the descriptor-based similarity is preferred over MFCC, and both are preferred over the random method. However, the detailed analysis shows

Score	Label
0-19	Very unnatural: repetitions, jumps, cuts.
20-39	Somewhat unnatural
40-59	Somewhat natural
60-79	Very natural
80-100	As natural as original

Table 2. Naturalness rating scale

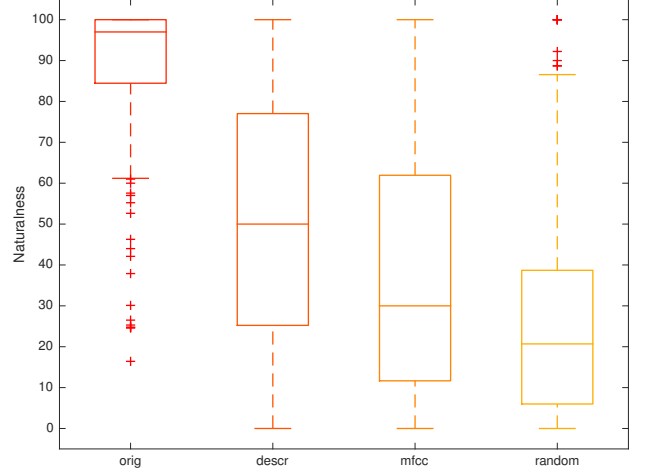


Figure 1. Box plot of the scaled naturalness ratings per type of stimulus, showing the median (middle line), quartile range (box), min/max (whiskers), and outliers (crosses).

that only for *Stadium Crowd* and *Water Faucet*, and to a lesser degree for *Formula One*, descriptor and MFCC-based distance are rated significantly different. We hypothesize that especially these sounds benefit greatly from the more detailed descriptors *Loudness* and *FundamentalFrequency*, as they contain sequences of pitched foreground events. For all sounds but *Traffic Jam* and *Baby Crying* the descriptor-based distance is significantly rated better than the random method, while for the MFCC-based distance this is only so for *Lapping Waves* and *Desert Wind*. Another remark is that for *Lapping Waves* and *Traffic Jam* the original is not rated significantly different from the descriptor-based method, and for the former sound this also applies to the MFCC-based method.

5. CONCLUSIONS AND FUTURE WORK

A possible explanation for the general superiority of the descriptor-based similarity measure over the MFCC based one is that the perceptual descriptors better capture certain aspects of the sound character of environmental textures beyond pure spectral shape (that is represented by MFCCs). We can hypothesise that some of this information is related to pitch content, as expressed by the *FundamentalFrequency* and *Noisiness* descriptors. More re-

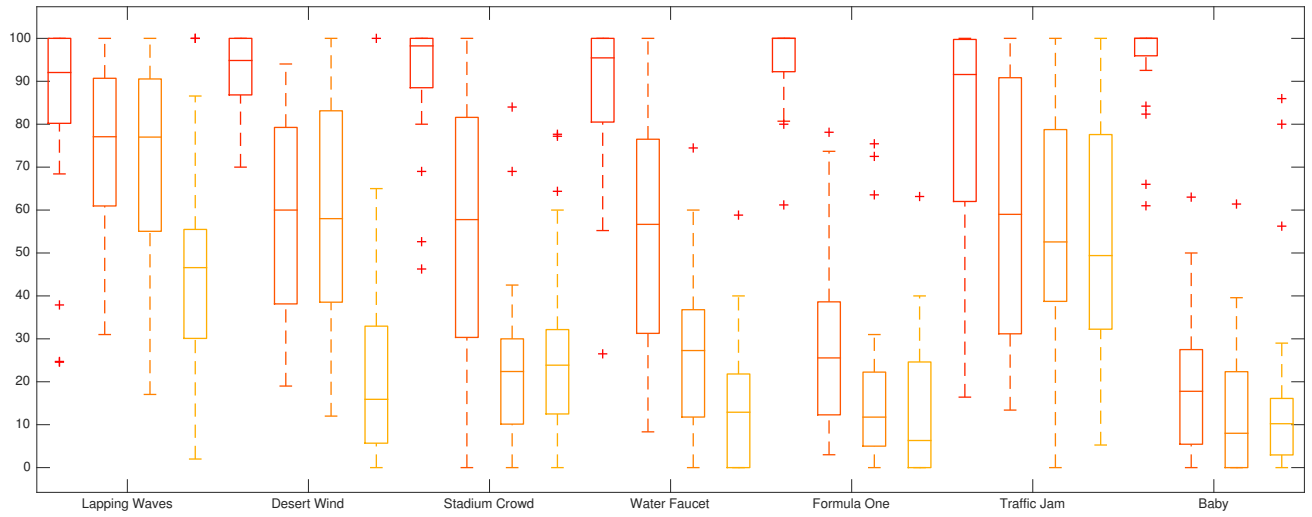


Figure 2. Box plot of the scaled naturalness ratings per source sound and type of stimulus (orig, descr, mfcc, random).

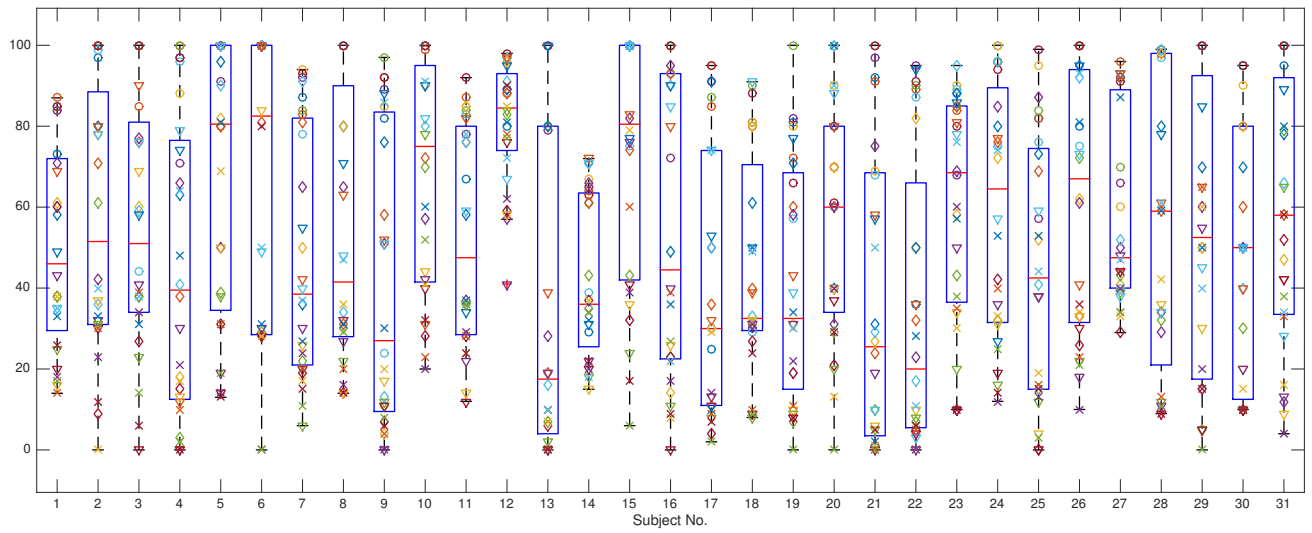


Figure 3. Box and dot plot of the per-subject naturalness rating prior to scaling (○ original, ◇ descr, ▽ mfcc, × random).

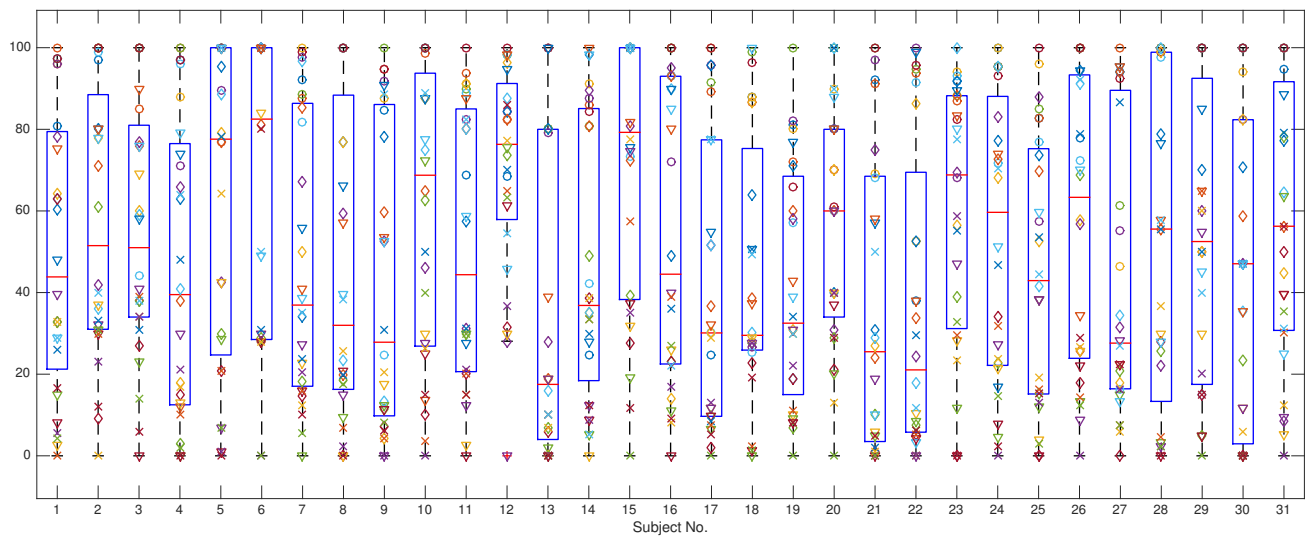


Figure 4. Box and dot plot of the per-subject naturalness rating after scaling (○ original, ◇ descr, ▽ mfcc, × random).

	orig descr	orig mfcc	orig random	descr mfcc	descr random	mfcc random
Lapping Waves	0.1772	0.2037	0.0000 ****	1.0000	0.0003 ***	0.0002 ***
Desert Wind	0.0000 ****	0.0000 ****	0.0000 ****	1.0000	0.0000 ****	0.0000 ****
Stadium Crowd	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	1.0000
Water Faucet	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.1179
Formula One	0.0000 ****	0.0000 ****	0.0000 ****	0.1576	0.0106 *	1.0000
Traffic Jam	0.0833	0.0338 *	0.0115 *	1.0000	1.0000	1.0000
Baby	0.0000 ****	0.0000 ****	0.0000 ****	1.0000	1.0000	1.0000
total	0.0000 ****	0.0000 ****	0.0000 ****	0.0002 ***	0.0000 ****	0.0002 ***

Table 4. P-values and significance class ⁵ for each pair of differences of means on unscaled naturalness ratings.

	orig descr	orig mfcc	orig random	descr mfcc	descr random	mfcc random
Lapping Waves	0.2360	0.1409	0.0000 ****	1.0000	0.0000 ****	0.0001 ***
Desert Wind	0.0000 ****	0.0000 ****	0.0000 ****	1.0000	0.0000 ****	0.0000 ****
Stadium Crowd	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	1.0000
Water Faucet	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.1408
Formula One	0.0000 ****	0.0000 ****	0.0000 ****	0.0365 *	0.0012 **	1.0000
Traffic Jam	0.0858	0.0297 *	0.0075 **	1.0000	1.0000	1.0000
Baby	0.0000 ****	0.0000 ****	0.0000 ****	1.0000	1.0000	1.0000
total	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0000 ****	0.0001 ****

Table 5. P-values and significance class ⁵ for each pair of differences of means on scaled naturalness ratings.

search is necessary to test this hypothesis and to link it with recent findings about fundamental mechanisms of sound texture perception [13].

While the presented method is not a sequence model that tries to model and generate the temporality of variation given an environmental recording, it can regenerate at least some of the naturally occurring variation in texture recordings. This has the two advantages of having a more varied output for background atmosphere sounds that uses the whole range of sound occurring in a source recording, and that the sound designer does not have to limit herself to stable textural recordings, or has to hunt down long-enough stretches of stable texture in longer recordings.

Although this is not the topic of this article, synthesis can be started off at specific-sounding grains in the recording as seeds, in order to start the texture with a given atmosphere (e.g. start with calm wind to not startle the listener at the beginning of a new scene with a gust of wind). This could, for instance, be achieved using a scatterplot interface that allows to visualise the timbral space in 2D as popularised by the CATART software ⁶. With a little future work, the texture could then be made to move towards another type of sound by specifying its feature vector and favouring transitions that move towards that point in the descriptor space. More future work should check the influ-

ence and possible automatic estimation of the neighbourhood parameter (the number of candidates c).

To conclude, we hope that this method can improve the workflow of sound designers for interactive or post-production applications, and further augment the advantages that procedural audio has to offer over fixed media in order to foster uptake by the industry.

Acknowledgments

The work presented here is partially funded by the French *Agence Nationale de la Recherche* (ANR) within the project *PHYSIS*, ANR-12-CORD-0006. The authors wish to thank Wei-Hsiang Liao for his groundwork on the online evaluation questionnaire, Axel Röbel, the *PHYSIS* project partners, and the Analysis–Synthesis and ISMM teams at Ircam.

References

- [1] Joan Bruna and Stéphane Mallat. Audio Texture Synthesis with Scattering Moments. page 5, November 2013. URL <http://arxiv.org/abs/1311.0407>.
- [2] Wim D’haes, Dirk van Dyck, and Xavier Rodet. PCA-based branch and bound search algorithms for computing K nearest neighbors. *Pattern Recognition Letters*, 24(9–10):1437–1451, 2003.
- [3] Shlomo Dubnov, Ziz Bar-Joseph, Ran El-Yaniv, Danny Lischinski, and Michael Werman. Synthesis

⁵ The significance level depending on the p-value is habitually represented by a number of stars as follows:

Level	*	**	***	****
$p \leq$	0.05	0.01	0.001	0.0001

⁶ <http://ismm.ircam.fr/catart>

of audio sound textures by learning and resampling of wavelet trees. *IEEE Computer Graphics and Applications*, 22(4):38–48, 2002.

- [4] Andy Farnell. *Designing Sound*. MIT Press, October 2010. ISBN 9780262014410. URL <http://mitpress.mit.edu/catalog/item/default.asp?tttype=2&tid=12282>.
- [5] M. Fröjd and A. Horner. Sound texture synthesis using an overlap-add/granular synthesis approach. *Journal of the Audio Engineering Society*, 57(1/2):29–37, 2009. URL <http://www.aes.org/e-lib/browse.cfm?elib=14805>.
- [6] Toni Heittola, Annamaria Mesaros, Dani Korpi, Antti Eronen, and Tuomas Virtanen. Method for creating location-specific audio textures. *EURASIP Journal on Audio, Speech and Music Processing*, 2014.
- [7] Stefan Kersten and Hendrik Purwins. Sound texture synthesis with hidden markov tree models in the wavelet domain. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, Barcelona, Spain, July 2010.
- [8] Anil Kokaram and Deirdre O’Regan. Wavelet based high resolution sound texture synthesis. In *Proceedings of the Audio Engineering Society Conference*, 6 2007. URL <http://www.aes.org/e-lib/browse.cfm?elib=13952>.
- [9] Wei-Hsiang Liao, Axel Roebel, and Wen-Yu Su. On the modeling of sound textures based on the STFT representation. In *16th Int. Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, September 2013. URL <http://architexte.ircam.fr/textes/Liao13a/>.
- [10] L. Lu, L. Wenying, and H.J. Zhang. Audio textures: Theory and applications. *IEEE Transactions on Speech and Audio Processing*, 12(2):156–167, 2004. ISSN 1063-6676. URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1284343.
- [11] 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)*, New Paltz, NY, October 18–21 2009.
- [12] Josh H McDermott and Eero P Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. *Neuron*, 71(5):926–40, September 2011. ISSN 1097-4199. doi: 10.1016/j.neuron.2011.06.032. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4143345&tool=pmcentrez&rendertype=abstract>.
- [13] Josh H McDermott, Michael Schemitsch, and Eero P Simoncelli. Summary statistics in auditory perception. *Nature neuroscience*, 16(4):493–8, April 2013. ISSN 1546-1726. doi: 10.1038/nn.3347. URL <http://www.ncbi.nlm.nih.gov/pubmed/23434915>.
- [14] Sean O’Leary and Axel Roebel. A two level montage approach to sound texture synthesis with treatment of unique events. In *DAFx*, Germany, September 2014. URL <http://architexte.ircam.fr/textes/OLeary14b/>.
- [15] Sean O’Leary and Axel Roebel. A montage approach to sound texture synthesis. In *EUSIPCO*, Lisbon, Portugal, September 2014. URL <http://architexte.ircam.fr/textes/OLeary14a/>.
- [16] Geoffroy Peeters. A large set of audio features for sound description (similarity and classification) in the Cuidado project. Technical Report version 1.0, Ircam – Centre Pompidou, Paris, France, April 2004. URL http://www.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf.
- [17] Norbert Schnell, Axel Röbel, Diemo Schwarz, Geoffroy Peeters, and Ricardo Borghesi. MuBu & friends – assembling tools for content based real-time interactive audio processing in Max/MSP. In *Proceedings of the International Computer Music Conference (ICMC)*, Montreal, Canada, August 2009.
- [18] Diemo Schwarz. Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24(2):92–104, March 2007. Special Section: Signal Processing for Sound Synthesis.
- [19] Diemo Schwarz. State of the art in sound texture synthesis. In *Proceedings of the COST-G6 Conference on Digital Audio Effects (DAFx)*, Paris, France, September 2011.
- [20] Diemo Schwarz and Baptiste Caramiaux. *Interactive Sound Texture Synthesis through Semi-Automatic User Annotations*. Lecture Notes in Computer Science. Springer International Publishing, 2014. doi: 10.1007/978-3-319-12976-1-23.
- [21] Diemo Schwarz and Norbert Schnell. Descriptor-based sound texture sampling. In *Proceedings of the International Conference on Sound and Music Computing (SMC)*, pages 510–515, Barcelona, Spain, July 2010.
- [22] Diemo Schwarz, Norbert Schnell, and Sebastien Guluni. Scalability in content-based navigation of sound databases. In *Proceedings of the International Computer Music Conference (ICMC)*, Montreal, QC, Canada, August 2009.