Parkkola, Kuura; Visakorpi, Vili; Wright, Alec

Analysis of Concatenative Synthesis Techniques

Published in:
Proceedings of the 2nd NordicSMC Conference

Published: 01/01/2021

Document Version
Publisher's PDF, also known as Version of record

ANALYSIS OF CONCATENATIVE SYNTHESIS TECHNIQUES

Kuura PARKKOLA[1], **Vili VISAKORPI**[2], and **Alec WRIGHT**[1]

[1]*Department of Signal Processing and Acoustics*, **Aalto University**, Espoo, Finland
[2]*Department of Applied Physics*, **Aalto University**, Espoo, Finland

ABSTRACT

Concatenative synthesis is a technique of Sound Texture Synthesis where a short clip of source audio is used to generate a continuous stream of similar sound. Several approaches for concatenative synthesis exist, but their output quality is highly dependent on the type and length of the source material, as well as the parameters used. The purpose of this paper is to test different synthesis techniques, parameters, and types of source material, to determine which algorithms best suit a variety of input sources. The study reveals that in most cases the complexity of the algorithm has a relatively small effect on the output quality. Noisy and repetitive sources produce the most natural sounding outputs even with short input clips and atom sizes; whereas, sources with a lot of variation over time typically require larger atom sizes and more source material to prevent audible artefacts in the synthesis output.

1. INTRODUCTION

Sound texture synthesis is a group of synthesis techniques aiming to generate arbitrarily long streams of audio mimicking real world soundscapes and sources. Sound texture synthesis spans a number of different approaches [1], one of which is *concatenative synthesis*: a subset of sound texture synthesis methods where a small amount of recorded sound is used to produce a sound texture similar to that of the original recording. The synthesis is performed through various looping techniques giving these algorithms an advantage in the production of natural and credible outputs, as much of the original audio remains intact. The field of sound texture synthesis is primarily driven by the movie, broadcast, and video game industries, where the flexibility provided by the family of techniques is well suited for media scenes of variable length.

The definition of a sound texture varies somewhat in previous work. Common definitions include stability on a large time-scale and possible inclusion of recurring micro events. A more deliberate definition is written in [2]. The precise definition is not regarded important for the results of this paper.

The human ear is exceptional at detecting recurring transients and sudden timbral changes in signals. The fundamental problem that needs to be solved by a decent concatenative synthesis algorithm is the randomization of the playback while retaining the naturalness of the source audio; the algorithm should introduce enough variation into the output stream to reduce noticeable periodicities while preserving a structural similarity to the original source material and avoiding discontinuities and sudden changes in the signal.

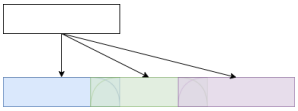


Figure 1. Behaviour of the looping synthesis technique.

This paper builds on the findings presented in [3]. The purpose of this study is to three concatenative synthesis techniques: looping, concatenation of randomized windows (RND) [4], and concatenative synthesis with descriptor controls (CSDC) [5], with a variety of source material to learn which approaches best suit the different source textures. Section 2 introduces the algorithms used in this paper along with their implementations. Section 3 makes observations about the general anatomy of sound textures. The methods used for the analysis of the algorithms and their outputs are covered in Section 4. Finally, the results are presented in Section 5 and the discussion and conclusions are in covered in Sections 6 and 7 respectively.

2. ALGORITHMS

This section covers the three algorithms evaluated in this paper in detail ordered from the simplest to the most advanced. Each algorithm is first briefly introduced and then the implementation and its parameters are explained. All of the algorithms were implemented with MATLAB. The implementations also utilize the Audio Toolbox.

2.1 Concatenative Synthesis with Looping

The most trivial method of extending a short clip of audio is to play it back over and over in an infinite loop. This approach makes no effort in disguising the repetition of the source material and with only elementary logic and few parameters, it is very quick to implement. The basic idea is depicted in Fig. 1.

The algorithm generates sound by copying the source clip into the output stream in a loop. The approach is described in close detail in Fig. 2. Crossfading is used to keep the signal power constant and to avoid discontinuities in the

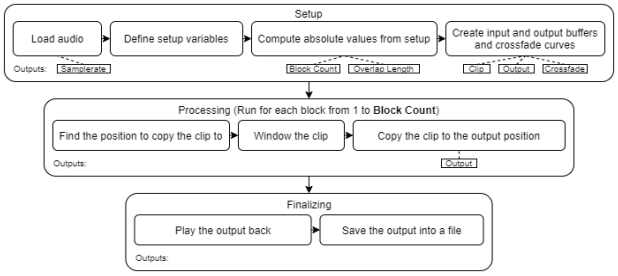


Figure 2. Structure of the looping synthesis implementation.

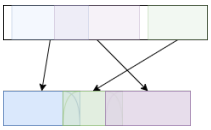


Figure 3. Behaviour of the randomized windowing technique.

output signal. The implementation uses equal power sinusoidal fades. The only parameters used by this algorithm are the lengths of the source material and the crossfade.

2.2 Concatenative Synthesis with Random Windows

The next step from looping synthesis is synthesis with randomized windows (RND) [4]. In this approach, the output stream is generated by randomly selecting sections from the source material and copying them to the output stream. Randomization hides the most obvious repetition in the output. A graphic depiction of the algorithm is presented in Fig. 3.

The windows used to generate the output stream are created by adding markers to random entry and exit positions in the source material. First the length of the window is defined at random from a given range. Then, an entry marker is randomly selected from the source material such that the distance from the marker to the end of the source clip is not less than the window length. Finally, the exit marker is added based on the entry marker and the window length and the material between the markers is copied into the output stream. The approach is described in further detail in Fig. 4.

The parameters used by this algorithm are the length of the source material, the range of window lengths, and the length of the crossfade.

2.3 Concatenative Synthesis with Descriptor Controls

In Concatenative Synthesis with Descriptor Controls (CSDC) the source material is divided into short windows called atoms [5]. To avoid sudden changes in the timbre of the output stream and thus preserving longer trends in the source material, the atoms are grouped by similarity based on a set of spectral descriptors. The selection of consecutive atoms is made according to the spectral similarity between the atoms. The idea is depicted in Fig. 5.

The implementation of CSDC here uses six descriptors: *spectral slope, spectral spread, spectral centroid, pitch, acoustic loudness and noisiness*, explained in further detail in [6]. The descriptors are used to give each atom a position in an n dimensional space where each dimension is described by one of the descriptors. The algorithm steps through the atoms by selecting the next atom for the output stream randomly from the nearest neighbours of the previously selected atom. The progression of the algorithm is described in Fig. 6.

When CSDC is used with relatively long atoms, the spectral content of each atom often shifts considerably over time. To combat this, the algorithm was extended with separate entry and exit descriptors. The extended CSDC (CSDCe) approach differs only in the selection of the spectral neighbors. Where the original CSDC implementation defines a single position for each atom in the spectral space, CSDCe defines a path with a start and end point. The nearest neighbors of one atom are selected from the nearest starting points to the end point of the atom in question.

Following the greater complexity of CSDC compared to its simpler counterparts, comes a wider range of tunable parameters. Where all of the algorithms presented in this paper accept parameters for the length of the source clip, size of the windows/atoms and the crossfade time between them, the results from CSDC are also affected by the weighting between the spectral descriptors. The weights of the descriptors warp the spectral space such that the higher the weight of a given descriptor is, the closer two atoms must be in its axis to be considered a near neighbour. The number of atoms to consider near neighbours also plays a role in the output quality: a smaller number of neighbors ensures greater similarity between the concurrent atoms but increases the likelihood of the algorithm getting stuck in a loop of very similar atoms.

3. ANATOMY OF SOUND TEXTURES

By analyzing different sound textures, we can observe that most of these textures can be classified along two axis: impulsiveness and repetitivity. Following these axis the textures can be grouped into four categories: repetitive and impulsive (Fig. 7a), repetitive and smooth (Fig. 7b), time-varying and impulsive (Fig. 7c), and time-varying and smooth (Fig. 7d). What should, however, be noted is that a sound texture can sometimes be formed of several different sounds, e.g. footsteps in a windy background, where the layered textures may have different characteristics. Some

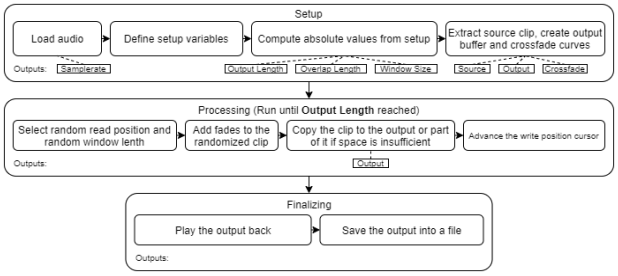


Figure 4. Structure of the randomized windowing implementation.

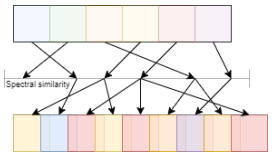


Figure 5. Behaviour of the CSDC technique.

sound textures consist mainly of static noise, these textures do not match the four groups being not repetitive nor impulsive but not varying in timbre over time either, thus requiring a class of their own. The sound of rain (Fig. 8) is one such static texture.

Repetitive sound textures are formed by very similar sound events equally spaced in time. The less these events evolve over time, the less there is information in the source material. Time varying sources are considerably more difficult to reproduce since there is very little natural repetition and thus, the textures contain more information. Impulsive sounds are somewhat easier to mimic since the sound events tend to be shorter than the atoms forming the synthesis, e.g. fire crackles. When the sound events are long, e.g. passing cars, they may get divided into multiple atoms which potentially leading to artifacts.

To hold enough information for natural texture reproduction, the source material should contain several complete time events. This is especially important for time varying textures where the synthesized signal should contain events that do not bear a close resemblance.

4. THE METHODS OF ANALYSIS

The synthesis results are evaluated according to both, a subjective aural analysis of the output stream and a more objective comparison between the spectrographic representation of the synthesized output and the original recording. A synthesis result is considered good if the spectrographic representation resembles the original recording over a long period, does not have noticeable repetitions and does not contain sudden timbral shifts. The aural analysis is expected to reveal shorter term anomalies not visible in the spectrograms.The evaluation was done by the authors and no formal listening tests were conducted.

Analysis is done for many different types of sound textures to find optimal algorithms and parameters for the four categories defined in Section 3. If multiple techniques provide similar results, the approach with the lowest complexity is preferred. The analyzed source materials are listed in Tab. 1.

Class	**Source**
Repetitive, impulsive	Punched card computer
	Drophammer
Repetitive, smooth	Ventilator
Time-varying, impulsive	Fireplace
	Footsteps
Time-varying, smooth	Applause
	Traffic
	Rain forest
Static	Rain
	Oil rig machinery

Table 1. The source material used in the analysis.

5. RESULTS

This section reports the findings from the testing and analysis of the synthesis techniques covered in Section 2. First, the general observations about the performance of the algorithms will be discussed followed by a closer inspection of the different types of sound textures. Both the impulsive and smooth repetitive sources showed such similar results that these texture classes were combined under a single subsection.

5.1 Overview of The Algorithms

In general, with an arbitrary choice of parameters and source material, the algorithms performed according to expectations. For the looping technique, one or two repetitions could be made without the recurrence becoming clearly noticeable. With atom sizes of more than 7 seconds, the periodicity started becoming less obvious. The RND approach managed to hide obvious repetition for roughly 5 times longer. The algorithm also managed to produce very natural outputs when given plenty of source material. Neither the CSDC nor the CSDCe approaches made much of an audible improvement beyond the RND technique; although, the longer scope results showed that the more advanced algorithms retained a greater structural similarity to the source material.

The properties of the probability distribution driving the RND solution led to the features near the center of the

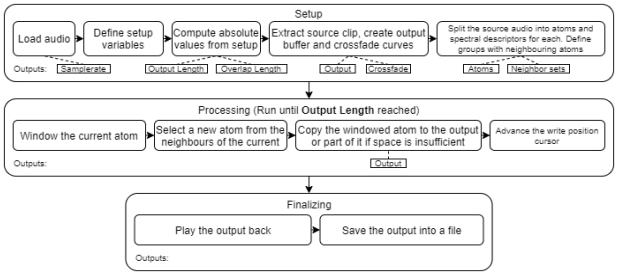
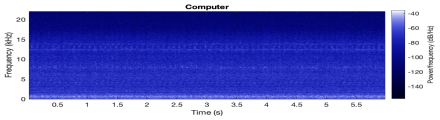
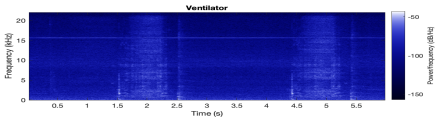


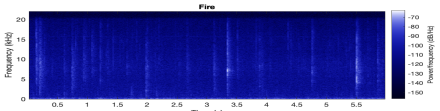
Figure 6. Structure of the CSDC implementation.



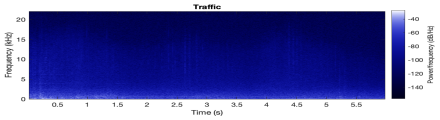
(a) Repetitive and impulsive texture of a punched card computer.



(b) Repetitive and smooth texture of a ventilator.



(c) Time-varying and impulsive texture of a fireplace.



(d) Time-varying and smooth texture of traffic.

Figure 7. Spectrogram representations of different types of sound textures.

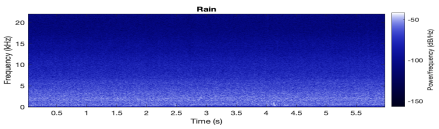
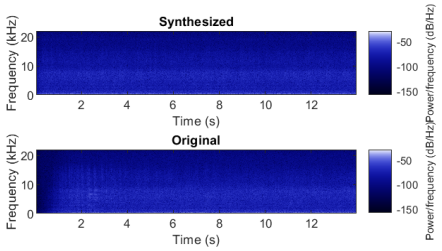


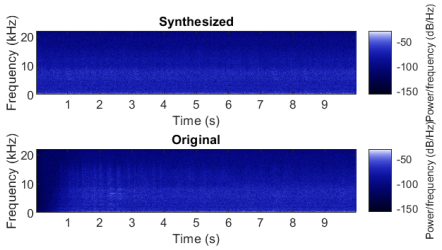
Figure 8. Static texture of rain.

source clip to be played back with a higher probability. With the clip length set to or near the length of the source material, events occurring near the middle could be heard repeating considerably more often. In some tests this behaviour made the implementation to perform worse than the looping approach.

Using simple crossfades to glue atoms together proved to be enough to remove noticeable artifacts from the output streams. Typically fades of around 45 ms were sufficent. With long fades the phase differences between the overlapping segments were occasionally noticeable. Although, this could likely be resolved using various signal processing techniques such as cross-correlation, the improvement



(a) Synthesized and recorded streams of a punched card computer.



(b) Synthesized and recorded streams of a ventilator.

Figure 9. Synthesized streams from static source material.

was considered beyond the scope of this study.

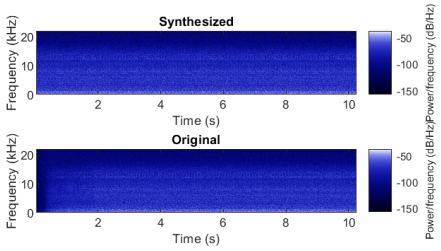
5.2 Static Textures

The static textures do not contain noticeable timbral shifts over time and thus do not need to leverage the spectrum matching capabilities of the CSDC algorithms. Both, the looping and RND approaches showed promise; however, the looping technique ended up producing better results when the length of the source audio was more than 5 seconds. The RND algorithm generated mostly excellent results, sometimes the same audio segments were played back in a quick succession, however, making the recurrences stand out. Phase mismatches in the output stream also became audible occasionally.

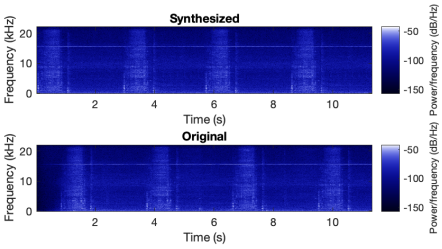
Evaluating the synthesis results according to their spectrograms (Fig. 9), the results are nearly identical. In the RND synthesis stream some barely noticeable artifacts can be identified for example around the 7 second mark.

5.3 Repetitive textures

Being inherently repetitive, the repetitive textures can be synthesized without trying to mask the recurrence in the



(a) Synthesized and recorded streams of a punched card computer.



(b) Synthesized and recorded streams of a ventilator.

Figure 10. Synthesized streams from repetitive source material.

output stream. With parameters synchronized with the period of the source material, the looping technique produced the best results. The strong periodicity in the source material managed to hide recurrence in the background noise of the synthesized streams. The spectrograms of both impulsive and smooth textures of a punched card computer (Fig. 10a) and a ventilator (Fig. 10b) respectively, show minimal differences between the artificial and the original textures.

With the repetitive textures the tonal characteristics of the source material varies little over time and thus, a review of the long term similarity between the synthesized signal and the original recording is not necessary.

5.4 Time-Varying Impulsive Textures

The RND approach provided the best results for source material with non-periodic short and impulsive sound events. The looping technique began sounding repetitive already at the first loop over even with a large amount of source audio. CSDC on the other hand refused to play the impulsive sound events altogether due to their abnormal timbre; instead, the algorithm only generated the background noise. The RND technique produced somewhat natural sounding outputs with even as little as 2 seconds of source audio with atom sizes of around 500 ms and 100 ms fades. Longer clips of source audio with more varying sound events produced even more realistic results. Increasing the atom sizes to 1 second and fade times to around 200 ms improved the quality of the background texture.

The spectrogram of a one minute stream rendered from 10 seconds of the sound of a fireplace is presented in Fig. 11. From these results it is obvious that the synthesized stream revisits the same sound events multiple times, structurally, however, the artificial stream has a very close resemblance to the original recording.

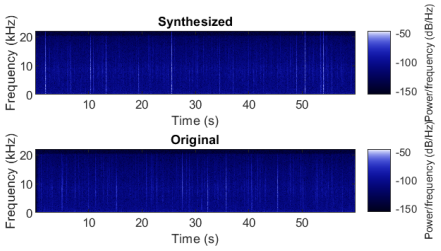


Figure 11. Synthesized and recorded streams of a fireplace.

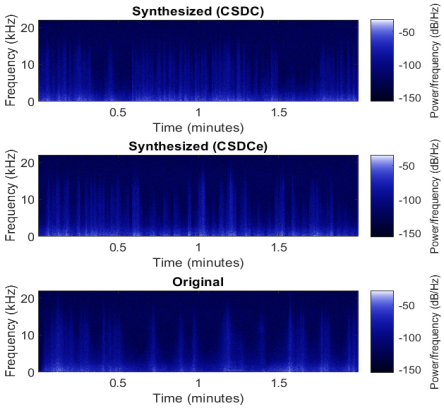


Figure 12. Synthesized and recorded streams of traffic.

5.5 Time-Varying Smooth Textures

Long evolving sound textures are the most complicated to reproduce. Although, the CSDC algorithm is made for these types of sound textures, the natural reproduction is still not trivial. A somewhat natural sounding synthesized stream requires a long clip of source audio and long atoms. The output quality of this implementation is yet not good enough to be considered applicable in practice.

Over a long period of time, the CSDC algoritms produce a great structural resemblance to the source material. This can be seen in Fig. 12 showing synthesis results generated from one minute of traffic noise with an atom length of 1.5 seconds, 200 ms fades and 15 spectrally closest atoms considered near neighbours. What is also immediately obvious is that the stream from the CSDCe approach with it's improved timbral flow bears a closer structural similarity to the original source than that of the basic CSDC implementation.

6. DISCUSSION

Most of the sound textures used as source material could be sufficiently well reproduced with an appropriate algorithm. For the repetitive textures, the looping technique left very little room for improvement. The reproduction of impulsive time-varying textures could be done with an acceptable quality, although, some clearly repeated transients did occur. The algorithm proposed in [7] shows great potential in adding variation to repeated transient sounds which

could improve the quality of synthesis of impulsive textures from less source material. The most noticeable artifacts in the CSDC output were caused by sound events being split over multiple atoms. This behaviour could likely be mitigated through the detection of such events to keep them intact. One technique with potential in achieving this is presented in [8]. The looping approach for the synthesis of static textures performed well; however, fine tuning the probability distribution and introducing phase matching for the RND algorithm would likely bring its quality to an equal level with less source material.

The examples given in this paper use unprocessed recordings from the real world, occasionally comprising complex mixtures of different types of textures. In practise, a mixture of sound textures could be created in a more controllable fashion by synthesizing the various textures on separate streams and blending them together to form more natural, multi-layered textures. This would also allow the use of an optimal technique for generating each texture. Similar results can also be achieved with synthesis approaches such as *montage synthesis* and *AudioTexture* presented in [3].

Although, progress in the field of sound texture synthesis has long passed the state of the art as it was a decade ago [1] and the focus in the field is shifting towards neural networks [9, 10], the results provided in this paper gives applicable insights for use in low-budget games and other media where the more complex modern approaches are difficult to implement.

7. CONCLUSIONS

This paper presented three common methods for concatenative sound texture synthesis and analyzed them according to their performance with varying source material. The three algorithms were first tested with arbitrary sources and parameters to gain information about the performance of the techniques. The source material was then divided into five classes in Section 3 and an optimal synthesis approach was chosen for each category.

With no particular attention to the parameters used, the looping technique generated unnatural and periodic output streams; the RND approach showed noticeable improvement on these results. The CSDC and CSDCe algorithms were not able to produce much of an audible improvement over the RND approach, though structurally the CSDC streams had a closer resemblance to the source material. Simple crossfading turned out to be a great way of reducing artifacts in the streams, some phase issues did, however, arise with long fades. The algorithms were not utilized with very short atom sizes like they were in [3] as this was found to produce less natural sounding outputs.

Repetitive sound textures with little evolution over time were best replicated with the looping technique when the length of the source material is an integer multiple of the period of the source. For time-varying textures with short time events such as clicks and cracks, RND produced very natural synthesized textures. CSDC achieved the best results with time-varying textures comprising long time events; however, the results were not good enough for application in practice. The static sound textures were most naturally reproduced with the looping approach, though RND seemed like a better option with minor improvements.

8. ENDNOTES

The source audio was obtained from the BBC sound effects library at `https://sound-effects.bbcrewind.co.uk/`. The algorithm implementations and sound examples covering their successes and shortcomings are available at `https://kuura.parkkola.fi/pub/material/concatenativesynthesis`.

9. REFERENCES

[1] D. Schwarz, "State of the art in sound texture synthesis," *Proceedings of the 14th International Conference on Digital Audio Effects, DAFx 2011*, Sep. 2011.

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

[3] D. Schwarz, A. Roebel, C. Yeh, and A. Laburthe, "Concatenative Sound Texture Synthesis Methods and Evaluation," pp. 217–224, Sep. 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01427393

[4] M. Fröjd and A. Horner, "Sound texture synthesis using an overlap-add/granular synthesis approach," *AES: Journal of the Audio Engineering Society*, vol. 57, pp. 29–37, 01 2009.

[5] D. Schwarz and S. O 'leary, "Smooth Granular Sound Texture Synthesis by Control of Timbral Similarity," p. 6, Jul. 2015. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01182793

[6] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Jan. 2004.

[7] J. Fagerström, S. Schlecht, and V. Välimäki, "One-to-many conversion for percussive samples," in *Proceedings of the 24th International Conference on Digital Audio Effects (DAFx20in21)*, 09 2021, pp. 129–135.

[8] A. Kumar, R. Singh, and B. Raj, "Detecting sound objects in audio recordings," in *2014 22nd European Signal Processing Conference (EUSIPCO)*, 2014, pp. 905–909.

[9] H. Caracalla and A. Roebel, "Sound texture synthesis using ri spectrograms," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 416–420.

[10] J. M. Antognini, M. Hoffman, and R. J. Weiss, "Audio texture synthesis with random neural networks: Improving diversity and quality," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3587–3591.