

SOUND TEXTURE SYNTHESIS USING CONVOLUTIONAL NEURAL NETWORKS

Hugo Caracalla

UMR STMS 9912

Sorbonne Université, IRCAM, CNRS,
Paris, France
hugo.caracalla@ircam.fr

Axel Roebel

UMR STMS 9912

IRCAM, Sorbonne Université, CNRS,
Paris, France
axel.roebel@ircam.fr

ABSTRACT

The following article introduces a new parametric synthesis algorithm for sound textures inspired by existing methods used for visual textures. Using a 2D Convolutional Neural Network (CNN), a sound signal is modified until the temporal cross-correlations of the feature maps of its log-spectrogram resemble those of a target texture. We show that the resulting synthesized sound signal is both different from the original and of high quality, while being able to reproduce singular events appearing in the original. This process is performed in the time domain, discarding the harmful phase recovery step which usually concludes synthesis performed in the time-frequency domain. It is also straightforward and flexible, as it does not require any fine tuning between several losses when synthesizing diverse sound textures. Synthesized spectrograms and sound signals are showcased, and a way of extending the synthesis in order to produce a sound of any length is also presented. We also discuss the choice of CNN, border effects in our synthesized signals and possible ways of modifying the algorithm in order to improve its current long computation time.

1. INTRODUCTION

The main difficulties encountered in sound texture synthesis become apparent when trying to properly define them. While examples of textures easily come to mind (e.g. environmental noises such as wind or rain, crowd hubbub, engine sounds, birds singing, etc.), pinpointing their common factors proves much harder: randomness appears to be one, along with a "background" aspect caused by an important number of indistinguishable small audio events happening at once. But this is not all, since we would still tend to call a sound including small occasional events happening in the foreground a texture. Hence, the definition offered by Saint-Arnaud [1], summed up by Schwartz [2], of "a superposition of small audio atoms overlapping randomly while following a higher level organization" is incomplete because it only encompasses completely stationary textures. It can even be argued that in reality no such texture can be observed: a synthesis algorithm strictly following this definition would thus be incomplete and of little practical use.

This means that a sound texture synthesis algorithm needs to be able to synthesize small indiscernible and random events but also singular, recognizable events, both harmonic (e.g. birds chirping) or not (e.g. crowd clapping). This extremely broad range of sounds is precisely what makes texture synthesis complex and

Copyright: © 2019 Hugo Caracalla et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

why common synthesis algorithms (for instance sinusoidal models) prove ill-suited for it, making it require dedicated ones.

Before presenting an overview of existing sound texture synthesis algorithm, we will first detail more precisely what is expected from them.

In the case of textures, "re-synthesis" is a term as fitting as "synthesis": starting from an existing texture, the goal is usually to create a sound that is different from the original while still being recognizable as the same kind of texture, as if it had been recorded only moments later. Although this is the prime goal of the algorithm, this obviously does not exclude the possibility of manipulating the synthesized texture. For instance, it could be desirable to allow the algorithm to synthesize texture lasting any arbitrary length of time, or to be able to have the synthesis evolve throughout time, altering its properties or progressively turning into another texture.

To achieve such a result, a broad variety of methods have been developed: for the needs of this article, we will split them into 3 different categories.

The first of those is *physics-based synthesis*. It regroups methods which goal is to first emulate the phenomenon at the source of the texture (for instance the impact of a drop of rain on a flat surface) via a physically informed model of it. From there one can simulate any number of events, dimensioning and randomizing them so as to fit the target texture. The result of this is a convincing physical simulation of the texture (see for instance [3]). While this method has the potential of being extremely controllable and allowing the manipulation of synthesis parameters that have a physical meaning, it also has the obvious flaw of not being flexible at all. Each algorithm correspond to one and only one kind of texture: a physical model of the rain will prove poorly suited to synthesize a flock of birds twittering.

The second is *granular synthesis*. It regroups methods in which the original texture is first chopped into milliseconds-long audio particles, then reordered and concatenated to reconstruct a new texture (see for instance [4]). While being quite versatile in the range of texture it is capable of re-synthesizing, such a method is also heavily dependent on the choice of atom size and requires more complex reordering methods when one tries to synthesize a broader array of textures. In particular, reconstructing any foreground event lasting more than the size of an atom proves a hard task.

The last synthesis category, to which our algorithm belongs, is *parametric synthesis*. It regroups synthesis methods in which the general goal is to establish a set of parameters to describe textures with. If those parameters are properly chosen, any two textures which parameters are equal will sound alike without necessarily being identical. From there, one would only need to modify a sound until its parameters are equal to those of a target texture

in order to re-synthesize it. While in theory this method is able to synthesize any and all kind of textures, in practice the quality and flexibility of the synthesis entirely depend on the choice of parameters. Those parameters need to describe the texture so as to hold enough information to re-synthesize a similar one, while not holding too much so as to not over-constraining the synthesis (in which case the only way of reaching a set of parameters would be by creating a perfect copy of the original texture). In addition to this, the parameters must be adapted to the description of the widest possible range of textures.

This paradigm is notably used by McDermott and Simoncelli in [5], where a set of statistics extracted from the critical and modulation bands are used as parameters. This algorithm gives convincing results for a broad range of textures, but fails when trying to synthesize textures containing singular events. Inspired by the work of Gatys et al. [6] who uses the cross-correlations between the feature maps of a trained 2D convolutional neural network (CNN) as parameters to synthesize visual textures, several attempts to convert this approach to audio have been made. In [7], Ulyanov and Lebedev use the same principle applied to spectrograms, with the frequency dimension acting as color depth and using a 1D CNN, to synthesize textures with moderate success. In [8], Antognini et al. add to this approach several constraints aimed at recreating rhythm with a better fidelity and increasing the diversity of results. While giving convincing results, this method requires fine tuning in order to balance the constraints. Since changing target texture implies tuning those synthesis parameters, this makes the algorithm lose in flexibility. In addition to this, both this method and Ulyanov and Lebedev's eventually output a spectrogram: it is then necessary to recover its phase and invert it to retrieve an audible time signal, using methods such as the Griffin and Lim algorithm [9]. This phase recovery step is an added burden to the synthesis as it tends to downgrade the quality of the audio signal, even more so when working with complex sounds such as textures.

Although not directly applied to sound texture synthesis, it may be noted that several works such as those of Grinstein et al. [10], Barry and Kim [11] and Tomczak [12] also use the same parametrization extracted from the feature maps of a CNN as part of audio style transfer processes.

In this work, we present a new parametric texture synthesis based on the method of Gatys et al.[6]. Our synthesis algorithm does not require any fine tuning or spectrogram inversion and works with a wide array of textures, including those presenting strong singular events. We also present a few examples to demonstrate its possibilities and proceed to discuss those results.

2. METHOD

Following the principle of parametric synthesis, we first define a set of parameters to represent a sound texture with.

2.1. Parametrization

2.1.1. Pre-processing

As our method is adapted from the work of Gatys et al. [6] on 2D images, we require a 2D representation of our sound signal.

To this effect we work with log-spectrograms. The log-spectrogram S is computed using the spectrogram X , taken as the magnitude of the short-term Fourier Transform (STFT) of the sound signal:

$$S = \frac{\log(1 + CX)}{\log(1 + C)} \quad (1)$$

where C a factor controlling compression: the larger C is, the more details we will get at low amplitudes. This choice of normalization is made to ensure that the spectrogram is both compressed by the log function and comprised between 0 and 1.

In practice we work with time signals sampled at 22050 Hz, and choose as STFT parameters a window length of 512 and hop-size of 256. The sound signals all have a length of 262400 samples so that their log-spectrograms are 1024 frames long, and have a bit depth of 16. The compression factor C is set to 1000.

For the rest of this article, any time-frequency matrix will have frequency as first axis, and time as second. For instance, $S(f, t)$ denotes the value of S at the f -th frequency bin and t -th time sample.

2.1.2. Network choice

Seeing as Gatys et al. use a network trained for image recognition, our first intuition was to use an equivalent network for working on spectrograms. As such we initially trained a simple deep 2D CNN on recordings taken from freesound.org in scene recognition in order to use it for synthesis. But in [13], Ustyuzhaninov et al. show that visual textures of the same quality as those obtained by Gatys et al. can be synthesized using a single-layer untrained CNN with various filter size instead of a trained CNN.

This proves to still hold for sound textures: the network we use to synthesize the textures presented in this article is a single-layer untrained 2D CNN using filters of different sizes and ReLU activation. Its single layer is made of 128 square filters of each of the sizes [3, 5, 7, 11, 15, 19, 23, 27] with a stride of (1, 1) and zero-padding so that the differently-sized convolutions can then be stacked, followed by the rectified linear unit (ReLU) activation layer. The weights from the filters are drawn from a uniform distribution between -0.05 and 0.05 , and no bias is applied.

For generalization's sake, the rest of the method is nonetheless presented with a network that has K layers (with K being potentially more than 1), although it is still valid when using a single-layer network.

It is worth noting that unlike the methods presented in [7, 8], we use the log-spectrograms as 2D images with time and frequency replacing the two space dimensions and not as a 1D signal with frequency as depth, hence the need for 2D convolutions.

2.1.3. Parameters computation

Let us denote $F_{i,(x,y)}^k$ the value of the i -th feature map of the k -th layer at the position (x, y) . In [6], Gatys et al. use the Gram matrices of each layer of the network as parameters. The (i, j) element of the gram matrix G^k of k -th Gram matrix is defined as the cross-correlation between the i -th and j -th feature maps of the layer:

$$G^k(i, j) = \sum_{(x,y)} F_i^k(x, y) F_j^k(x, y) \quad (2)$$

The parameter set is chosen as the list of gram matrices from G^1 to G^K . Although this proves a good choice for visual textures, such parameters cannot be directly used in the case of sound

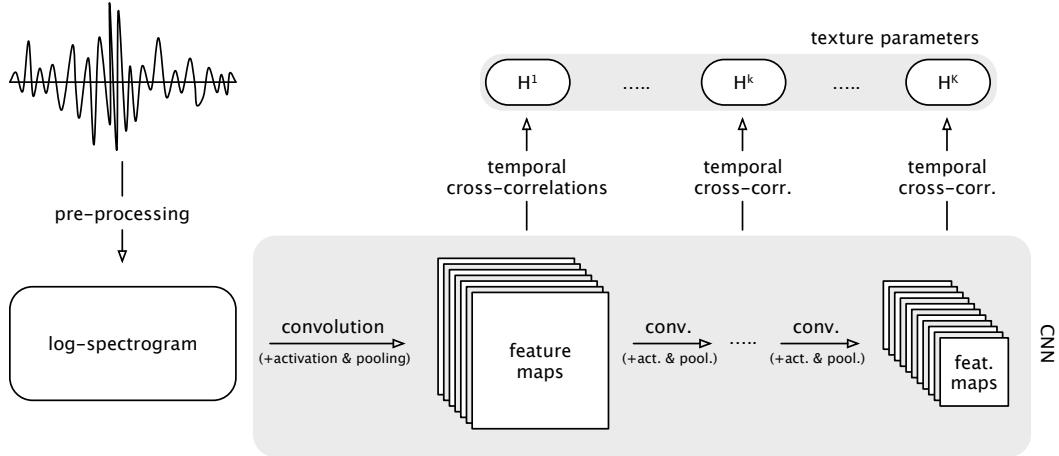


Figure 1: Computation of the 3-dimensional parameter tensors from the temporal cross-correlations of the feature maps of a CNN, .

textures. Indeed, those Gram matrices average all spatial information when performing a sum over all positions (x, y) , thus implying that the parametrization should be invariant in both directions. This does not translate well to sound, since sound textures behave differently regarding time and regarding frequency: while we wish for a pseudo-stationarity over the time dimension, there is no reason for there to be any invariance to pitch-shifting. As such, we instead use the 3-dimensional tensors H^k defined as:

$$H^k(i, j, x) = \sum_y F_i^k(x, y) F_j^k(x, y) \quad (3)$$

Defined this way, the tensors H^k with $k \in [1, K]$ that we use as parameters do indeed average all information from the time dimension, but keep the information regarding the frequency dimension intact, thus achieving our goal. The parameter extraction process is represented in Figure 1.

2.2. Texture loss

As mentioned in Section 1, the main goal of parametric synthesis is to create a sound which has the same parameters values as a those of a target sound. Seeing as we now have a parameter set, we only need to define a quantitative error function which will then be minimized throughout the synthesis process. To that effect, we use a simple distance function between the two sets similarly to Ustyuzhaninov et al. [13]:

$$\mathcal{L} = \sum_k \frac{\|\tilde{H}^k - H^k\|_2}{\|\tilde{H}^k\|_2} \quad (4)$$

with $\|\cdot\|_2$ denoting the L2 norm, and the tilde denoting the target texture parameters.

2.3. Optimization

The last step of the synthesis process is to create a sound signal which minimizes the texture loss. Since the chain of operations leading to the computation of the texture loss is differentiable (despite passing through the complex domain due to the STFT: see [14] for further insight), we may use any optimization algorithm requiring the gradient of the error function to iteratively modify a

sound until it reaches a satisfying minimum of the loss function (similar time domain synthesis can be found in [11, 12]).

We observed that performing the optimization on the log-spectrogram is iteration-wise faster than performing the optimization directly on the sound signal (both of them being initialized using white noises). This seems to indicate that the texture loss is simpler to minimize when working in the time-frequency domain rather than when working directly in the time domain. To take advantage of that fact, we first perform a quick synthesis of a log-spectrogram and invert it using a random phase matrix (which would correspond to performing one step of the Griffin-Lim algorithm): while this inversion raises the value of the texture loss, it still makes for a good initialization of the optimization in the time domain. This allows us to skip a major part of the optimization process on the sound signal. Once performed this optimization results in a sound signal which minimizes the texture loss, meaning its parameters values are close to those of the target texture. The synthesis process is illustrated in Figure 2.

As in [6] and [8], we found that the L-BFGS algorithm (introduced in [15]) converges fast and yields good audio results. Starting from a white noise image, we perform 1000 iterations of it in the time-frequency domain to create the initialization of the time domain optimization. The time domain optimization is then performed over 10000 iterations. Using a GeForce GTX 1080 Ti GPU, the whole process takes around an hour.

3. EXPERIMENTS

3.1. Experimental results

The log-spectrograms of both target and synthesized textures are shown in Figure 3 for three sounds: a wildlife scene with crickets chirping in the background and a bird singing in the foreground (recognizable to its inverted "v"-shaped patterns), birds singing both in the background and in the foreground (with one strongly standing out the mid-frequency range), and the hubbub of a crowd chatting. The audio signals of all three are available for listening online¹ along with other textures such as wind, bees and fire sounds.

¹See <http://recherche.ircam.fr/anasynt/caracalla/dafx19/synthesis/>

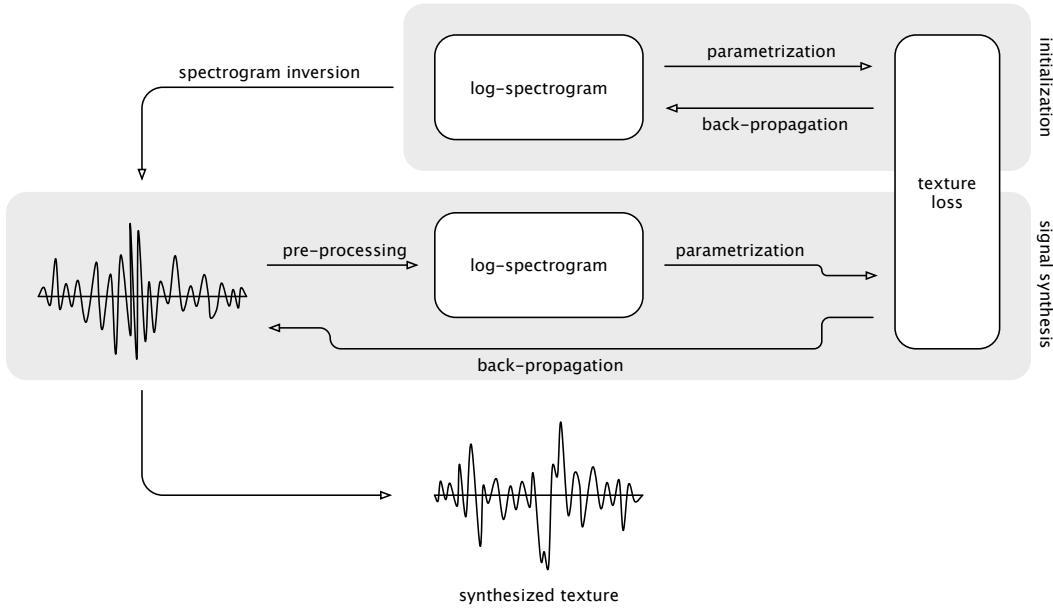


Figure 2: *Organization of the texture synthesis algorithm: a quick optimization on a log-spectrogram is performed in the time-frequency domain and its result is inverted. It then serves as initialization for the main optimization performed on a sound signal.*

3.2. Discussion

3.2.1. Results analysis

The strength of this texture synthesis algorithm lies in the fact that it manages to both synthesize background and foreground events convincingly. The crowd chatter, which includes no singular events, is as well recreated as the bird singing loudly in front of a flock.

Another advantage to it is the absence of any parameter tuning: unlike the method presented by Antognini et al. [8] where three losses need to be balanced through the optimization process and potentially from one texture to another, the texture loss used here is straightforward and requires no tuning. This way the algorithm can effectively be used without needing to take into account which texture is being synthesized.

In addition to this, because the final optimization is performed on the time signal our algorithm does not end with a spectrogram inversion (the harm a phase retrieval process could bring is clearly noticeable when inverting spectrograms of existing texture and comparing the resulting signal to the original).

3.2.2. Untrained vs. trained network

As mentioned in Section 2.1.2, both trained and untrained network have been used in visual texture synthesis with success. The main argument in [6] is that the CNN trained for image recognition has learned filters adapted to common shapes encountered in images. This, coupled with the depth of the network, is what supposedly allows the network to recreate a large array of shapes when trying to reproduce the cross-correlations between the activations of its filters once the network has been fed an image (in this case, a visual texture). This argumentation is challenged by Ustyuzhaninov et al. in [13], who demonstrate that a single-layered untrained CNN can perform as well as a trained network given enough filters. This would tend to imply that given enough random filters, the space of

shapes recreated when synchronizing some of those filters is wide enough to compensate for the lack of training.

This translates seamlessly to sound textures: while we first worked with networks trained for acoustics scene recognition in an attempt to emulate the process of [6], experiments with untrained network show that they perform just as well. This being said, it could be interesting to explore the difference between the use of the two further: for instance, trained CNN might require less filters than untrained ones, thus making our parameter tensors lighter and the computations faster. The depth of the trained CNN might also help it capture correlations across distant events in the spectrogram. This would indeed be useful, seeing as birds textures from Figure 3 show that while the algorithm manages to reproduce the local pattern of bird cries well, it fails to reproduce the larger pattern of groups of cries separated by gaps of a few seconds. This is also quite audible when listening to the attempt at (non-texture) singing voice synthesis: since the human voice is rich in harmonics, it spans over a large portion of the frequency spectrum. Because our algorithm does not enforce long-distance correlations, the upper harmonics are not synchronized with the lower ones, thus creating another high-pitched voice speaking on its own. In the fire synthesis, this is also clearly noticeable when looking at impacts: since those short and sharp events span over most frequencies, the algorithm has trouble generating them and mostly manages to recreate impacts that only span over part of the frequency axis, resulting in less convincing synthesized textures. This could potentially be solved even when using untrained network by choosing bigger filters, which would then "see" larger chunks of the log-spectrogram at once.

3.2.3. Border effect

Our texture synthesis presents one intriguing property: at the start and end of the sound, the synthesized texture is identical to the original one (for instance, this is slightly visible at the start of the

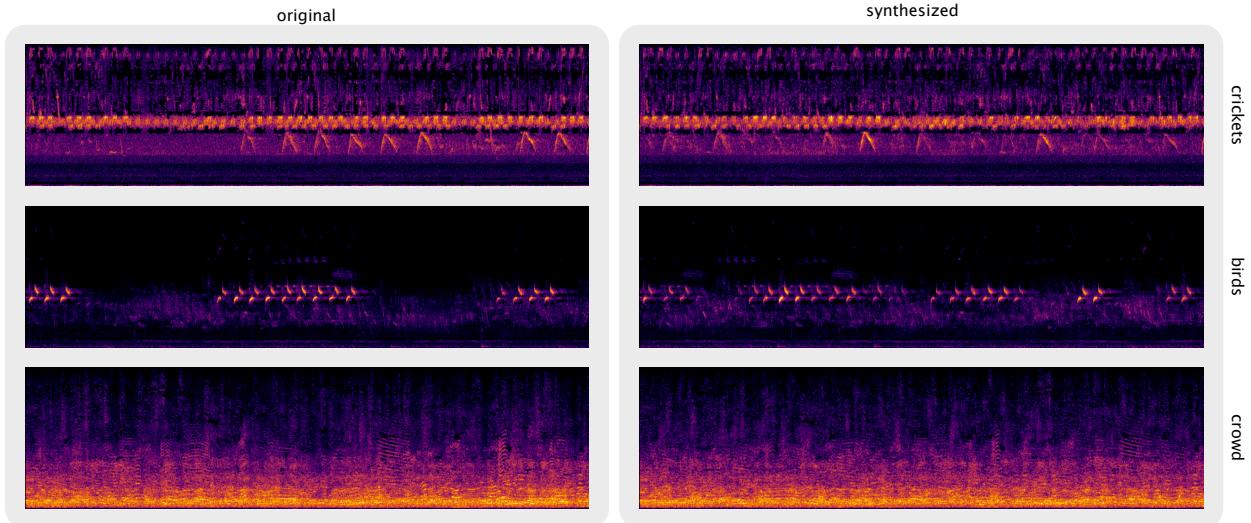


Figure 3: Log-spectrograms of both originals textures (on the left) and synthesized textures (on the right) using the method presented in this article.

log-spectrogram of the birds texture from Figure 3). In all of our synthesis, the leftmost and rightmost frames present exactly the same patterns in both original and synthesized textures. Although this effect is interesting, it dissipates quickly and only affects the time dimension: this means that even if we were to not get rid of this artifact, slightly cropping the start and end of the synthesized texture does completely discard it.

Gatys et al. [6] notice a similar effect in his visual texture synthesis where a distinctive part of the image is always reproduced around the same spot, and suggest that this effect originated from the zero-padding used in the convolution layers. To test this theory we experimented with synthesizing textures using only "valid" convolutions (i.e. without any padding): our results still presented the same border artifacts, which would indicate that they do not originate from the zero-padding. We do not have any alternative explanation to present at the moment.

It is worth noticing that this effect lasts around the length of the biggest filter used (in our case, 27 frames): for now, this means that we need to choose a filter size large enough to ensure the good reproduction of correlations between events, while small enough so that the border effect does not spoil too much of the interior of the synthesized texture.

3.2.4. Computation time

As mentioned in Section 2.1.3, computation time is for now far behind real-time since it takes around an hour to synthesize 12 seconds of audio with one GPU. While tedious for now, this process could potentially be alleviated by removing as much redundant information from the target parameters as possible (as of now, we have around $\sim 100M$ parameters in the parameter tensors H_k).

Using the same network as described in Section 2.1.2, we removed all cross-correlations between filters of different size from the parameter tensor: this dropped the number of parameters to $\sim 16M$ without altering the quality of synthesized textures. Another lead is to use a trained CNN instead of an untrained one, seeing as trained filters should prove efficient at describing pat-

terns without needing to be as numerous as in an untrained CNN. We believe it should also be possible to drop the number of parameters even lower, for instance by using principal component analysis to select which cross-correlations need to be imposed over the synthesized signal as Gatys et al. [6] did.

Another potential lightening of the algorithm could come from changing signal representation: so far we have used log-spectrograms, but using another time-frequency representation could be greatly beneficial. For instance, using mel bands instead of the raw frequency bins of the spectrogram would reduce the number of parameters while staying perceptually sensible.

3.2.5. Extension

The basis for the texture synthesis having been set, it is also possible to develop on it and add ways of creatively alter the synthesized texture. An example of manipulation is the creation of an indefinitely long sound texture.

In order to do so, we use a principle resembling the "exquisite cadaver" game, where one has to continue the drawing of someone else while only seeing the border of the other's drawing. In our case we first synthesize an initial sound texture from a given target and copy the end of it onto the start of a white noise signal: we then perform another synthesis using this signal as initialization and keeping the same target texture. While doing this, we also prevent the optimization to be performed on the section that was copied from the previous synthesis. This results in a continuous texture seamlessly extending on the copied part, thus being able to perfectly follow where the previous synthesis left off. We only need to concatenate the newly generated texture to the previous one to create a longer sound texture. This process can obviously be repeated any number of times so as to obtain a texture of any desired length. One iteration of this process is shown on figure 4, while an example of such a synthesis is available online.²

²See http://recherche.ircam.fr/anasynt/caracalla/dafx19/extended_synthesis/

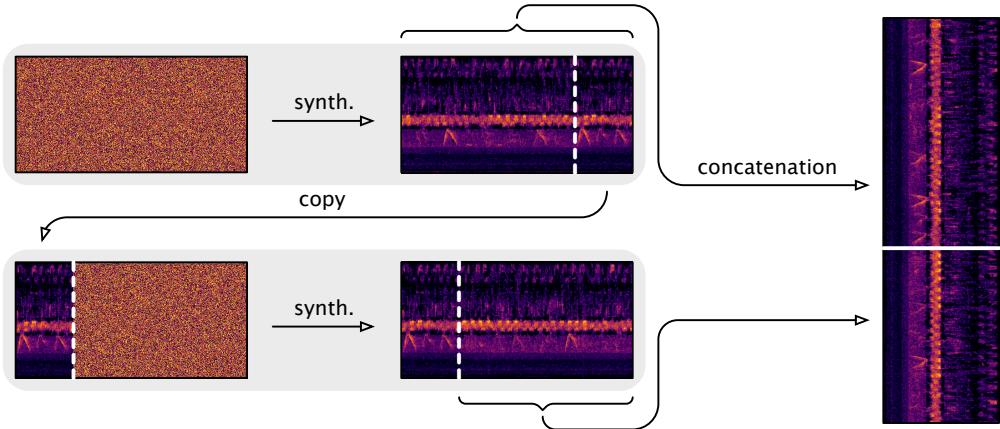


Figure 4: Extension of a first synthesis by copying its end onto the start of a white noise signal. This signal is then used as initialization of another synthesis, while preventing the common part from being modified. The tiles are then concatenated to form a longer texture. Sound signals are represented by their log-spectrogram for explanation's sake.

4. CONCLUSIONS

We introduced a new parametric texture synthesis based on the work of Gatys et al. [6] in visual textures: using the temporal cross-correlations between the feature maps of a CNN as parameters, we iteratively modify a sound signal until the values of its parameters reach those of a target texture. While the input of the CNN is the log-spectrogram of the sound, the optimization process is made directly in the time domain so as to avoid any phase recovery step in the synthesis.

The algorithm yields convincing results on a wide array of texture, even if they include singular events in the foreground. It can be straightforwardly applied without requiring the fine tuning of synthesis parameters from one texture to another. Its major flaws as of now lie in its long computation time and its trouble re-synthesizing correlations of events far apart in the log-spectrogram. A number of possible ways to address the first issue have been presented, for instance by subsampling the parameters tensor and altering the time-frequency representation. As for the second, the influence of the CNN architecture, and most notably the shape of its filters, are currently being investigated.

5. REFERENCES

- [1] Nicolas Saint-Arnaud, *Classification of sound textures*, Ph.D. thesis, Massachusetts Institute of Technology, 1995.
- [2] Diemo Schwarz, “State of the art in sound texture synthesis,” in *Digital Audio Effects (DAFx)*, 2011, pp. 221–232.
- [3] James F. O’Brien, Chen Shen, and Christine M. Gatchalian, “Synthesizing sounds from rigid-body simulations,” in *ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2002, pp. 175–181.
- [4] Diemo Schwarz, “Corpus-based concatenative synthesis,” *IEEE signal processing magazine*, vol. 24, no. 2, pp. 92–104, 2007.
- [5] Josh H McDermott and Eero P Simoncelli, “Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis,” *Neuron*, vol. 71, no. 5, pp. 926–940, 2011.
- [6] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge, “Texture synthesis using convolutional neural networks,” in *Advances in neural information processing systems*, 2015, pp. 262–270.
- [7] Dmitry Ulyanov and Vadim Lebedev, “Audio texture synthesis and style transfer,” Available at <https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/>, 2016.
- [8] Joseph Antognini, Matt Hoffman, and Ron J. Weiss, “Synthesizing diverse, high-quality audio textures,” *arXiv preprint arXiv:1806.08002*, 2018.
- [9] Daniel Griffin and Jae Lim, “Signal estimation from modified short-time fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.
- [10] Eric Grinstein, Ngoc Q.K. Duong, Alexey Ozerov, and Patrick Pérez, “Audio style transfer,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 586–590.
- [11] Shaun Barry and Youngmoo Kim, “Style transfer for musical audio using multiple time-frequency representations,” 2018, Available at <https://openreview.net/?id=BybQ7zWCb>.
- [12] Maciek Tomczak, Carl Southall, and Jason Hockman, “Audio style transfer with rhythmic constraints,” in *Digital Audio Effects (DAFx)*, 2018, pp. 45–50.
- [13] Ivan Ustyuzhaninov, Wieland Brendel, Leon A. Gatys, and Matthias Bethge, “Texture synthesis using shallow convolutional networks with random filters,” *arXiv preprint arXiv:1606.00021*, 2016.
- [14] Hugo Caracalla and Axel Roebel, “Gradient conversion between time and frequency domains using wirtinger calculus,” in *Digital Audio Effects (DAFx)*, 2017, pp. 234–238.
- [15] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal, “Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization,” *ACM Transactions on Mathematical Software (TOMS)*, vol. 23, no. 4, pp. 550–560, 1997.