MOSAIC STYLE TRANSFER USING SPARSE AUTOCORRELOGRAMS

Dan MacKinlay and Zdravko Botev
School of Mathematics and Statistics, UNSW Sydney
dan@danmackinlay.name

ABSTRACT

We introduce a novel mosaic synthesis algorithm for musical style transfer using the autocorrelogram as a feature map. We decompose the autocorrelogram feature map sparsely in a decaying sinusoid basis, using that decomposition as an interpolation scheme in feature space. This efficiently provides gradient information in the mosaicing optimization, including gradients of the challenging time-scale parameters, which are usually computationally intractable for discretely sampled signals. The required calculations are straightforward to parallelize on vector-processing hardware. Our implementation of the method provides good quality output and novel musical effects in example tasks by itself and can also be integrated into alternative mosaicing methods.

1. INTRODUCTION

Mosaicing synthesis is a particular approach to the style transfer problem. As with all style transfer methods, the goal is combining two signals, a source, and a target, to produce a hybrid output signal with qualities of each, which we call a *mosaic*. A musical application of these methods would typically use the 'style' of one signal, the timbre, to express the 'content' of another, a melody. Concretely, if the target were a trumpet playing a melody, and the source a recording of a singing vocalist, the mosaic might aim to emulate the vocalist singing that melody.

There exist a variety of problem definitions of, and associated algorithms for, mosaicing synthesis; e.g. [8, 12, 13, 22, 23, 35, 41, 45], partially summarized in [32]. In mosaicing specifically, we accomplish style transfer using a dictionary-based granular synthesis method, which constructs its output by superposition of transformed short recordings, *grains*, from an audio dictionary, in the time or, more recently, spectral domains [1,7,16].

The granular synthesis methods in themselves are well understood and widely deployed in industrial applications. They comprise a significant proportion of the music industry market for software synthesizers, are integrated into every major Digital Audio Workstation package, and have been extensively researched – see e.g. [31] and references therein.

The extension of granular synthesis into a style-transfer problem as mosaicing is less well-understood. In this setting we choose the parameters of a granular synthesis so as to optimally approximate a desired target audio signal in the sense of optimising some measure of acoustic similarity. Typically, this implies approximating, in the sense of minimising some approximation loss, the power spectral density (PSD) of the target signal. Applications for this include musical accompaniment, creative musical effects, or user customization of speech synthesis [10].

Our sparse autocorrelogram method advances the capabilities of musical mosaicing applications, by leveraging a feature map that is related to, but more convenient than, classical PSD methods. This method is enabled by two major innovations.

Firstly, we define signal similarity through the *autocorrelogram*, a representation of the signal as covariance with delayed versions of itself. The autocorrelogram and its relationship to PSD is well-known (e.g. [44]) but our use in mosaicing synthesis appears novel. Although we use the autocorrelogram in a standalone procedure, it may be included in the feature vectors of loss functions of other mosaic techniques and is thus of independent interest.

Secondly, we decompose the high-dimensional empirical autocorrelogram into a sparse dictionary of decaying sinusoids. By interpolating discrete signals, this procedure calculates both error and gradients efficiently, enabling gradient-based optimization. The resulting technique is flexible and straightforward to parallelize on modern Single Instruction Multiple Data (SIMD) architectures such as Graphics Processing Units (GPUs).

We make our Python code [1] openly available for public use. We thereby aim to facilitate both the investigations of future researchers and the immediate application of these methods by musicians. Comparisons are made with benchmark mosaicing implementation, NiMFKS [7].

2. PRIOR WORK

Style transfer techniques, construed broadly, have a long history in signal processing research. Early work in this area begins with the channel vocoder [17], via various innovations to the modern repertoire of methods which includes innovations such as neural style transfer methods

Attribution: Dan MacKinlay and Zdravko Botev . "Mosaic style transfer using sparse autocorrelograms", 20th International Society for Music Information Retrieval Conference, Delft, The Netherlands, 2019.

[1] https://github.com/danmackinlay/mosaicing_omp_ismir_2019/

[19, 21, 43]. In the style transfer field, the mosaicing techniques form a sub field which fix a choice of synthesis to dictionary-based granular synthesis techniques.

We are concerned with the musical applications of style transfer. The archetypal task in this context is using the timbre of the 'style' signal to express the melodic 'content' of another. Concretely, if the target were a trumpet playing a melody, and the source a recording of a singing vocalist, the output should emulate the vocalist singing that melody.

In mosaicing synthesis, the task of choosing synthesis parameters to produce the desired output is non-trivial and subject of ongoing interest. Notable recent progress includes matrix factorization methods to decompose audio [1, 7, 16], various improvements in spectral matrix factorization [1, 7, 16] and optimization over features [8, 11, 36]. However, few methods can conveniently handle time-scaling of audio, so that time-scale parameters must be ignored, or selected by exhaustive search. One recent exception is Sound Retiler [1], which claims to handle time shifting via tensor decomposition. It is in this area that we make our main contribution, by the application of autocorrelogram features in this task.

While the autocorrelogram itself is not new in audio synthesis (e.g. [38]), our application to the mosaicing problem seems novel. The autocorrelogram-based analysis in combination with sparse coding induces a novel and analytically differentiable expression for the time scale parameter, and it is this we use to solve the mosaic problems.

3. PROBLEM DESCRIPTION

3.1 Audio signals and notation

We work with audio signals, a Hilbert space $\mathcal{H}$ of real L_2 functions $f : \mathbb{R} \to \mathbb{R}$ mapping time to instantaneous signal pressure level. Where the argument of the signal is clear, we abbreviate notation, writing for example, $t \mapsto f(at)$ as $f(at)$. We will handle transforms on signals $f(.)$ such as the autocorrelogram $\mathcal{A}$, and Fourier transform $\mathcal{F}$. Where not clear from context which argument of the signal with respect to which the transform is taken, we indicate it with a subscript to the transform. Thus $\mathcal{F}_t\{f(s,t)\}(\xi) := \int e^{-2\pi i t \xi} f(s,t)\mathrm{d}t$. Where we specify a weight v for the inner product or norm, we write it as a subscript, i.e. $\langle f, g\rangle_v := \int_{\mathbb{R}} v(t)f(t)g(t)\mathrm{d}t$.

In practice we do not observe continuous audio signals, but discretely sampled observations of signals. Sampling fidelity will be assumed, requiring signals are band-limited to some suitably low cutoff period Ω. We scale time so that the sample period $T = 1$ and $\Omega > 1/2$. The sampling process is a train of Dirac impulses, and inner products with a discrete signals are defined

$$\langle g, f\rangle_v := \sum_{t\in\mathbb{Z}} v(t)g(t)f(t). \tag{1}$$

We denote length-M vectors in bold, $\boldsymbol{x} = [x_1, x_2, \ldots, x_M]^\intercal$.

3.2 Mosaicing

Given a target signal f_0, we seek an approximant, the mosaic $\hat{f}_0$, as a sparse linear combination of scaled signals, called *codes*, from a source *dictionary* $\mathfrak{G} := \{g_1, \ldots, g_D\}$ subject to a maximum budget of J codes. In our earlier style transfer example, say, f_0 would be the recorded trumpet melody and $\mathfrak{G}$, recordings of the singing vocalist. For a fixed dictionary the mosaic is specified completely by the length-J parameter vectors $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\rho}$ and written

$$\hat{f}_0(t; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\rho}) = \sum_{j=1}^{J} \alpha_j g_{\gamma_j}(\rho_j t). \tag{2}$$

The problem requires selecting approximately optimal values for parameter vectors

$$\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\rho}\} \simeq \operatorname*{argmin}_{\{\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\rho}\}} d\left(\hat{f}_0(t; \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\rho}),\, f_0(t)\right), \tag{3}$$

where $\rho_j \in \mathbb{R}^+, \alpha_j \in \mathbb{R}, \gamma_j \in \{1, \ldots, D\}$ and $d : \mathcal{H} \times \mathcal{H} \mapsto \mathbb{R}^+$ is a distance function quantifying the poorness of the approximation. In contrast to sparse coding for signal compression, $\hat{f}_0$ is an intentionally imperfect approximation of f_0, possessing qualities of both the source and target signals, hence the designation style *transfer*.

4. AUTOCORRELATION MOSAICING METHOD

The autocorrelogram mosaicing method has two stages.

1. In the pre-training stage, autocorrelogram features are computed from the source signals, and decomposed in a dictionary of decaying sinusoids.
2. In the inference stage, we search our dictionary of autocorrelogram decompositions for matches to the autocorrelogram of the target signal, and solve an inverse problem, synthesizing a corresponding mosaic from our result.

Both stages leverage convenient properties of autocorrelograms, and sparse dictionary decompositions, which we now introduce.

4.1 Properties of autocorrelograms

We now motivate the use of the autocorrelogram in our feature map. As with other style transfer methods we face the challenge that sample values of a time domain audio signal f are only indirectly indicative of how human listeners will perceive it. For audio analysis, one typically operates on a feature map $\mathcal{P}\{f\}$ which is in some sense closer to human perception of these signals. Specifically, we aim to find a feature map such that two signals are similar if some distance between their feature vectors is small, i.e. the similarity of f and $\hat{f}$ is high iff the distance $d_{\mathcal{P}}(f, \hat{f}) := \|\mathcal{P}\{f_0\} - \mathcal{P}\{\hat{f}\}\|$ is low, with some choice of norm $\|\cdot\|$. We would like $d_{\mathcal{P}}$ to approximate specifically *psychoacoustic* similarity, which is to say $d_{\mathcal{P}}(f, \hat{f})$ is small iff a typical human listener would perceive f and $\hat{f}$

as similar. Ideally the image of the feature map should also be of lower dimension than f, and $d_{\mathcal{F}}$ should be computationally efficient to manipulate.

True psychoacoustic similarity is not well-defined, so practical algorithms settle for feature maps compromising between convenience and psychoacoustic plausibility. Usually, feature maps are empirical PSDs [8, 23], or are derived from the PSD, as with the Mel-Frequency Cepstral Coefficient (MFCC) [27] or the Constant-Q transform [6]. These maps induce expensive mosaicing optimization problems [11, 12]. MFCCs for example, are suitable for low-dimensional indexing and search but are hard to invert. A raw empirical PSD is easier to invert, via, e.g. Griffin-Lim iteration, but of the same order of dimensionality as the original signal and thus difficult to search. One could ameliorate this difficulty if a computationally convenient feature map could be found which was well-behaved under operations of scaling and superposition, as in Eq. 2, so that one could conduct as much calculation as possible in the feature space.

These desiderata suggest the autocorrelogram map

$$\mathcal{A}\{f\} :\xi \mapsto (\xi \mapsto \langle f(t), f(t-\xi)\rangle)\,. \tag{4}$$

This is the deterministic covariance between $f(t)$ and $f(t-\xi)$. The autocorrelogram is an even function in ξ, so we work with one-sided autocorrelograms $\mathbb{R}^+ \to \mathbb{R}$. Autocorrelogram-like transforms are implicated in the neurological processing of harmonic audio by human listeners [3, 9, 25, 26]. For our purposes, the supposed neurological basis is a secondary consideration to the demonstrated empirical usefulness in psychoacoustic tasks, most notably in pitch-detection [30, 37, 40]. In this regard it resembles the cepstral analysis method [5], which also effectively identifies small numbers of periodic components by analysing a pointwise non-linear transformation of the power spectrogram , but unlike the cepstrum it is well-behaved under superposition.

Specifically, brief calculation shows the following useful properties: a) Multiplication by a constant $c \in \mathbb{R}$:

$$\mathcal{A}\{cf\}(\xi) = c^2\mathcal{A}\{f\}(\xi). \tag{5}$$

b) Time scaling:

$$\mathcal{A}\{f(rt)\}(\xi) = \frac{1}{r}\mathcal{A}\{f\}\left(\frac{\xi}{r}\right) \tag{6}$$

c) Randomized addition:

$$\mathbb{E}\left[\mathcal{A}\{S_1 f + S_2 f'\}(\xi)\right] = \mathcal{A}\{f\}(\xi) + \mathcal{A}\{f'\}(\xi), \tag{7}$$

where $\{S_i\}$ are i.i.d. Rademacher variables, taking values in $\{+1, -1\}$ with equal probability.

We note two obstacles to the application of these formulae in the mosaicing problem. Firstly, Eq. 6 is not well-defined for the discrete signals that comprise the usual subject matter of digital signal processing. We will handle discrete signals by continuous interpolants, which turn out to be practically sufficient approximations. Secondly, the additive rule c) is valid only in expectation, via the contrivance of introducing Rademacher random variables. Solving for the deterministic case by accounting for phase cancellation is indeed possible, but considerably more involved, and constitutes an active area of research in its own right in, e.g. the Overlap-Add [15, 42], and phase retrieval [24, 34] literatures. As the randomised solution also turns out in practice to be already sufficient for many tasks, we defer such extensions to future work.

In order to construct these interpolants efficiently, we decompose discrete autocorrelograms using a matching pursuit, which we now introduce.

4.2 Orthogonal matching pursuit

In orthogonal matching pursuit (OMP) [14, 28], given a target signal f_0 and a dictionary of *code* signals $\mathfrak{D} = \{g_\theta\}_{\theta\in\Theta}$, one finds a decomposition $\hat{f}_0 = \mathrm{OMP}_{\mathfrak{D},K}(f_0)$ of form

$$f_0 \simeq \underset{\mathfrak{D},K}{\mathrm{OMP}}(f_0) := \sum_{i=1}^{K} \mu_i g_{\theta_i}. \tag{8}$$

A solution is a parameter vector $[\theta_1, \dots, \theta_K] \in \Theta^k$ and code weights $[\mu_1, \dots, \mu_K] \in \mathbb{R}^K$ which nearly minimize $\|f_0 - \hat{f}_0\|$. We require that f_0 and all codes g_θ are L_2 integrable and not null, i.e. possessing positive norm, $\|g_\theta\| > 0$.

The OMP algorithm is as follows.

1. Initialization. Let the first residual be $r_0 := f$. Set step counter $k \leftarrow 1$.
2. Find θ_k such that (possibly approximately)
$$\theta_k = \underset{\theta}{\mathrm{argmax}}\, A(r_k, g_\theta) \tag{9}$$
where A is the *normalized code product*
$$A(r_k, g_\theta) := \frac{\langle r_{k-1}, g_\theta\rangle}{\|g_\theta\|}. \tag{10}$$
3. Solve the least sum of squares problem
$$[\mu_1^k, \dots \mu_k^k] = \underset{[\mu_1,\dots,\mu_k]}{\mathrm{argmin}} \Big\| \sum_{1\leq\ell\leq k} \mu_\ell g_{\theta_\ell} - f_0 \Big\| \tag{11}$$
giving kth decomposition $\hat{f}^k = \sum_{1\leq\ell\leq k} \mu_\ell g_{\theta_\ell}$.
4. Update the residual $r_{k+1} = f_0 - \hat{f}^k$.
5. If $k = K$, stop, otherwise set $k \leftarrow k+1$ and repeat from step (2).

We allow the components of θ to be either a) a discrete and finite, or b) a continuous parameter. For finitely enumerable components $\theta_{\text{finite}} \subseteq \theta$ we maximize normalized code product in Eq. 9 by enumeration. For continuous components $\theta_{\text{cts}} \subseteq \theta$ we assume that we can choose θ_{cts}

approximately by iterative optimization using the gradient $\nabla_{\theta_{\text{cts}}} A(r_k, g_\theta)$. As the objective may not attain a global maximum, we choose $I \geq 1$ different initial guesses, and select the best local optimum attained. A first order gradient ascent with fixed number of steps performs well in our examples and moreover requires no branching instructions, as suits our goal of a SIMD-compatible algorithm.

4.3 Sparse approximate autocorrelograms

In the pre-training stage, we find autocorrelograms for each of the empirical source autocorrelogram codes in $\mathfrak{G}$, decomposing them into a dictionary of sparse OMP matches, $\mathfrak{M}$. It is this dictionary which we search for mosaic matches, using matches here to identify approximately matching codes in the original space $\mathfrak{G}$.

In this section we use ξ as the free argument for signals, and restrict $\xi > 0$. For the interpolant dictionary we use decaying sinusoids

$$\mathfrak{S} := \{h(\xi; \omega, \tau, \phi) := \cos(\omega\xi + \phi)e^{-\tau\xi} : \phi, \tau, \omega \in \mathbb{R}\}. \tag{12}$$

The dictionary choice must ultimately be justified by empirical performance, which we demonstrate in the final section of the paper. It is notable that there are also *a priori* reasons for favouring this one for musical audio. Firstly, this basis will decompose an autocorrelogram into a global approximant, rather than a piecewise interpolant, as with for example polynomial splines. Evaluations of such an interpolant are tractable to parallelise without branching instructions, and therefore better suited to modern SIMD architectures.

Secondly, decaying sinusoid models are effective in compactly decomposing time-domain audio [20], and the nature of the autocorrelogram suggests that they could be similarly useful and even more compact in decomposing autocorrelograms. The space of superpositions of decaying sinusoids is, by inspection, closed under the autocorrelogram transform, so it is at just as plausible to represent autocorrelograms in a such a decaying sinusoid dictionary. The question remains how compact such a representation is. Analytic expansion of the superposition of many decaying sinusoids is a lengthy exercise in elementary calculus. However, we have reason to suspect that the amplitude coefficient of most terms in such expansions will negligible. Recall the Wiener-Khintchine theorem, which says that, for signals of finite energy, assuming all these terms are well-defined,

$$\mathcal{F}_\xi\{\mathcal{A}\{f\}(\xi)\}(s) = |\mathcal{F}_t\{f(t)\}(s)|^2$$

where $\mathcal{F}_\xi\{f(\xi)\}$ is the Fourier transform of signal $\xi \mapsto f(\xi)$. This tells us that the magnitude of sinusoidal components of the autocorrelogram are squared with respect to the magnitude of sinusoidal components of the PSD, and thus relatively sparser. This indicates that for autocorrelograms of musical signals, which are well approximated by a superposition of sinusoidal signals, the autocorrelogram could often be approximated with comparable relative error by a yet smaller number of sinusoidal signals, as can be seen in Fig. 1. Moreover, we know that the envelope of musical audio spectral content decays eventually super-exponentially with frequency [18] and thus high frequency content of an autocorrelogram will in general be proportionally even lower. This latter fact additionally implies that the autocorrelogram calculations might even be down-sampled with little loss in information content, and some computational saving.

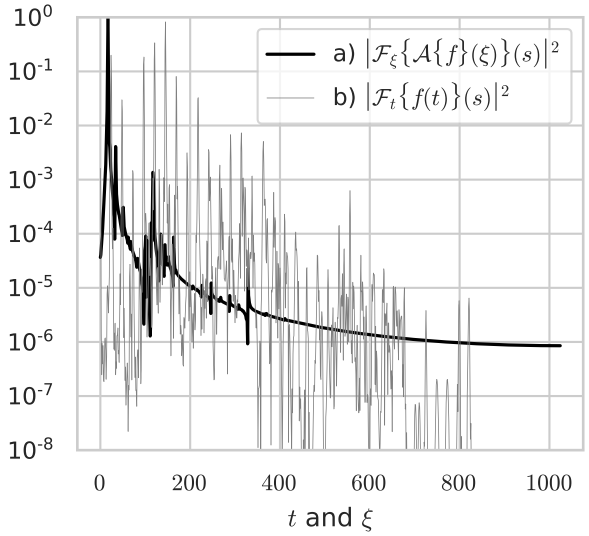

Figure 1. The relatively simple form of (a) the PSD of the autocorrelogram versus (b) the PSD of the signal itself. Signal is a length 2048 recording of a trumpet note onset. The scale of the vertical axis is arbitrary, and signals have been normalized for comparison. Sample period is $1/44100s$.

Implementing the decomposition is straightforward. For each code $g \in \mathfrak{G}$ we perform the following calculation: First, we find the empirical autocorrelogram $\mathcal{A}\{g\}$ at L points $\xi = 0, 1, \ldots, (L-1)$ with Eq. 4.

Next, we decompose each $\hat{G} = \text{OMP}_{\mathfrak{S},C}(\mathcal{A}\{g\})$ over the decaying sinusoid dictionary, as defined in Eq. 12. There are many methods of fitting decaying sinusoids to time series [2, 29, 33], but OMP is convenient in the current application [20] as we may re-use the same algorithm in the reconstruction stage of this algorithm. Autocorrelograms of musical audio in our experiments are highly sparse with respect to this decaying sinusoid dictionary, typically achieving negligible residual error with number of components $C \leq 4$.

We will apply the OMP with product $\langle \cdot, \cdot \rangle_v$ weighted by $v(\xi) := \mathbb{I}\{[0, L)\}(\xi)/L$, returning parameters $\{\tau_i, \omega_i, \phi_i\}$ and code weights μ_i. We first find the normalized code product (Eq. 10) in closed form. Substituting in Eq. 12 gives

$$A(r(\xi), h(\xi; \omega, \tau, \phi)\}) = \frac{\langle r_i(\xi), \cos(\omega\xi + \phi)e^{\tau\xi}\rangle_v}{\|\cos(\omega\xi + \phi)e^{\tau\xi}\|_v}. \tag{13}$$

The numerator is simply Eq. 1. Applying Euler identities gives the denominator

$$\begin{aligned} &\|\cos(\omega\xi + \phi)e^{-\tau\xi}\|_v^2 \\ &= \frac{1}{2}\int_0^L (1 + \cos(2\omega\xi + 2\phi))e^{-2\tau\xi}\mathrm{d}\xi \\ &= \tfrac{e^{-2\xi\tau}}{2}\tfrac{(\omega\sin(2\xi\omega+2\phi)-\tau\cos(2\xi\omega+2\phi))}{4\tau^2+4\omega^2}\Big|_{\xi=0}^{\xi=L} + \tfrac{1-e^{-2L\tau}}{4\tau} \end{aligned} \tag{14}$$

Combining Eq. 1 and Eq. 14 gives a closed form normalized code product (Eq. 13), from which we can explicitly calculate gradients in τ, ω, ϕ as desired. Note that although the original signal is discrete, our decomposition is a continuous near-interpolant for it.

From these decompositions we construct the dictionary

$$\mathfrak{M} := \{\hat{G}_\gamma(\rho\xi) : \gamma \in (1, \dots, D), \rho \in \mathbb{R}^+\}. \tag{15}$$

4.4 Synthesizing the mosaic

In the second, inference, stage we construct a mosaic $\hat{f}_0$ given a target f_0. Here we match the discrete autocorrelogram $F_0 := \mathcal{A}\{f_0\}$ by a second OMP decomposition $\hat{F}_0 := \mathrm{OMP}_{\mathfrak{M},J}(F_0)$, into

$$\hat{F}_0(\xi) := \sum_{j=1}^{J} \kappa_j \hat{G}_{\gamma_j}(\rho_j\xi) \tag{16}$$

for index parameters $\{\gamma_i, \rho_i\}$ and weights κ_i. The OMP has already been introduced, but we pause to verify that it may be applied to this new context. Since each $\hat{G}_{\gamma_j}$ is a linear combination of decaying sinusoids (Eq. 12), the normalizing denominator of the code product (Eq. 10) is again a linear combination of decaying sinusoids, so its integral has a (lengthy) closed form as a linear combination of integrals (Eq. 14), and we can find an explicit gradient $\nabla_\rho A(r_k, \rho)$. Thus we may find $\hat{F}_0$ as required.

Now we wish to construct $\hat{f}_0$ (Eq. 2) such that

$$\mathbb{E}[\mathcal{A}\{\hat{f}_0\}] = \hat{F}_0. \tag{17}$$

Choosing $\hat{f}_0 := \sum_j S_j \alpha_j g_{\gamma_j}(\rho_j t)$ by matching pursuit, simulating S_j independent Rademacher variates, and applying Eqns. 5, 6, 7 to Eq. 2, we find

$$\begin{aligned} \mathbb{E}[\mathcal{A}\{\hat{f}_0\}] &= \mathbb{E}\left[\mathcal{A}\left\{\sum_j S_j \alpha_j g_{\gamma_j}(\rho_j t)\right\}(\xi)\right] \\ &= \sum_j \frac{\alpha_j^2}{\rho_j}\mathcal{A}\left\{g_{\gamma_j}(t)\right\}(\rho_j\xi) \\ &\simeq \sum_j \frac{\alpha_j^2}{\rho_j}\hat{G}_{\gamma_j}(\rho_j\xi). \end{aligned} \tag{18}$$

By inspection,

$$\alpha_j = S_j\sqrt{|\rho_j||\kappa_j|} \tag{19}$$

satisfies Eq. 17. We resample the original discrete dictionary codes to target time scale ρ_i by band-limited sinc interpolation [39]. Finally, we substitute the resulting α_j into Eq. 2 and superpose grains to realize the desired mosaic.

4.5 Localized matching

So far we have discussed entire signals, implicitly assuming them to be brief. The autocorrelogram, taken globally over a long signal such as an entire musical piece, no longer estimates the local, stylistic characteristics. Just as one adapts the discrete Fourier transform for long signals into the Short-Time Fourier Transform (STFT) [4], so do we adapt the autocorrelogram mosaic method, applying it locally. A simple localization is to slice signals into short frames of fixed duration M, which are called *grains* by convention. As in the STFT, we multiply each frame point-wise with real window function w, supported on $[0, M]$ with $\|w\| = 1$. Hereafter, we assume a *sine window*, $w(t) := 2\sin(\pi t/M)\mathbb{I}[0, M]/M$. We fix hop length $H < M$. Next, we localize $\mathfrak{G}$ into a new dictionary whose codes are precisely these time-shifted grains (disallowing zero-energy grains).

$$\mathfrak{G}^{w,H} := \{w(t)g(t - \phi) : g \in \mathfrak{G}, \phi/H \in \mathbb{Z}, \|g'\| > 0\}. \tag{20}$$

In musical material a localized dictionary tends to high redundancy and marginal return on search effort decreases. Rather than proceeding exhaustively, we keep the search tractable by searching a pseudorandom subset of fixed size, where the size of this pseudorandom subset is a user selectable parameter.

In the synthesis stage, we localize the target signal, $f_0^w(t; \phi) := w(t)f_0(t - \phi)$, constructing a local mosaic $\hat{f}_0^w(t; \phi)$ from $\mathfrak{G}^{w,H}$ for $\phi = 0, H, 2H, \dots$ Finally, we superpose the local mosaics into a global one,

$$\hat{f}_0(t) = \sum_{\ell\in\mathbb{Z}} \hat{f}_0^w(t + H\ell; H\ell). \tag{21}$$

5. EXPERIMENTS

As an initial example we transfer style with target f_0 trumpet solo[2] and source audio a vocal recording.[3] Audio is sampled with a period of 1/44100s. We fix $M = 8192$, $H = M/2$, $L = 1024$, $C = 4$, $J = 1$, $I = 12$ and reasonable default parameters for the optimization routines. Examining the spectrogram Fig. 2 illustrates phenomena compatible with our claims: In the mosaic we observe local features of the source with the larger structure of the target, to wit, the pitch contours of the trumpet solo with a spectral distribution somewhat like the human voice.

[2] credit Mihai Sorohan

[3] credit Emm Collins

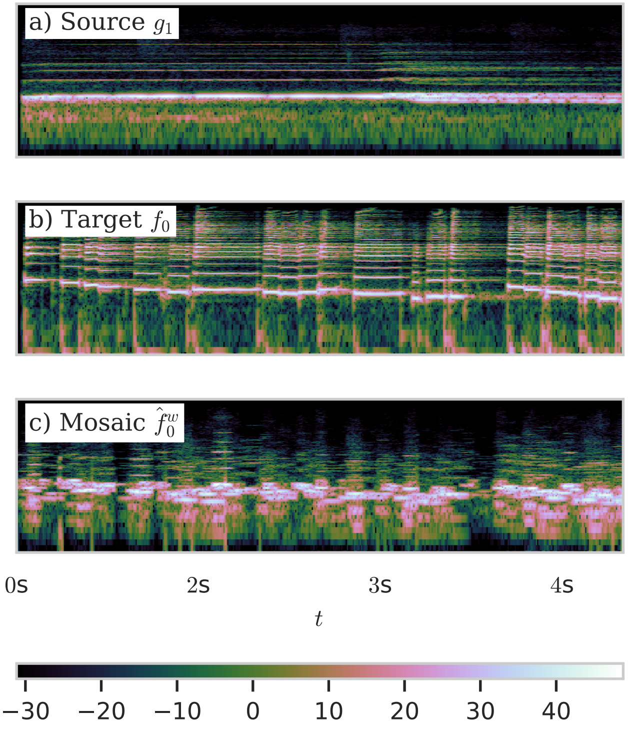

Figure 2. Power spectral density of signals a) source vocal recording b) target trumpet recording and c) resulting mosaic. Frequency increases up vertical axis, intensity in dB with arbitrary normalization.

We next apply the algorithm across a small corpus and compare our results against the mosaicing algorithm NiMFKS [7].[4] NiMFKS is a useful benchmark for mosaicing synthesis, incorporating many different user-selectable loss functions and decompositions methods from elsewhere in the literature, and possessing openly available code.[5] Their method generalises classical mosaicing by using a non negative PSD factorization to further decompose grains into a sparse product of activations and responses. Unlike our method it does not infer optimal time scaling of audio.

Performance evaluation of mosaicing methods is subjective. In the following, we will nevertheless attempt to describe the behaviors of the two algorithms as objectively as we are able. In order to challenge the NiMFKS model, our corpus samples are tuned to a variety of different root notes, scales and audio ranges, including Indonesian, western and centerless tunings. Style transfer is applied to every pairing of samples. Parameters are left at default values in each algorithm. These may be heard in the supplemental material. Subjectively, neither method seems to produce naturalistic outputs for all pairs of source and target audio. NiMFKS seems ascendant where the source audio is polyphonic and the factorization succeeds at decomposing different notes where our method cannot. On the other hand, where the target tuning is not spanned by the source, the sparse autocorrelogram method is able to produce smoother and better related mosaics by transposing source grains to match the target. Occasionally the sparse autocorrelogram mosaics sound rough during rapid articulations; the method could possibly be improved in these cases by adaptive selection of grain size, or tuning of the free hyperparameters in the model, or extension with non-randomised reconstruction methods. Even in these cases, however, simultaneous playback of the target and the mosaic reveals that we maintain harmonic relationships with the target audio. As such, even this imperfect reconstruction can be regarded as an exotic musical effect. In summary, even at this early stage, our method succeeds in extending mosaic methods to previously intractable tasks, and produces musically interesting output.

6. CONCLUSION

By combining autocorrelogram feature maps and interpolating matching pursuit, we have extended the library of methods of audio mosaicing style transfer. Our method in isolation produces interesting results on the sample data with little tuning. Work remains to be done in analysing the robustness and generality of the method, and selecting optimal tradeoff of cost and quality of different style transfer tasks under different choices of user parameters. More work also remains to be done in integrating this method with existing ones. The flexible loss function of, for example, NiMFKS could be augmented to include autocorrelogram features, and the autocorrelogram approach can be applied to spectrally decomposed signals, which are still audio signals. However, the ease with which we produce good results suggests that further extensions and refinements are worthy of pursuit.

7. ACKNOWLEDGEMENTS

We are indebted to Graham Coleman, James Nichols, the Dadabots and MTF Stockholm for illuminating conversations that greatly improved this work.

8. REFERENCES

[1] H. F. Aarabi and G. Peeters. Music Retiler: Using NMF2D Source Separation for Audio Mosaicing. In *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, AM'18, pages 27:1–27:7, New York, NY, USA, 2018. ACM.

[2] H. Barkhuijsen, R. de Beer, W. M. J. Bovée, and D. van Ormondt. Retrieval of frequencies, amplitudes, damping factors, and phases from time-domain signals using a linear least-squares procedure. *Journal of Magnetic Resonance (1969)*, 61(3):465–481, Feb. 1985.

[3] G. M. Bidelman and A. Krishnan. Neural correlates of consonance, dissonance, and the hierarchy of musical

[4] It would be instructive to compare against mosaicing method Music Retiler [1], which claims to handle time scaling of audio via a different method, should the source code become available.

[5] `https://code.soundsoftware.ac.uk/projects/nimfks`

pitch in the human brainstem. *Journal of Neuroscience*, 29(42):13165–13171, Oct. 2009.

[4] R. B. Blackman and J. W. Tukey. *The Measurement of Power Spectra from the Point of View of Communications Engineering*. Dover Publications, New York, 1959.

[5] B. P. Bogert, M. J. R. Healy, and J. W. Tukey. The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. In *Symposium on Time Series Analysis*, pages 209–243, 1963.

[6] J. C. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, Jan. 1991.

[7] M. Buch, E. Quinton, and B. L. Sturm. NichtnegativeMatrixFaktorisierungnutzendesKlangsynthesenSystem (NiMFKS): Extensions of NMF-based concatenative sound synthesis. In *Proceedings of the 20th International Conference on Digital Audio Effects*, page 7, Edinburgh, 2017.

[8] M. Caetano and X. Rodet. Musical Instrument Sound Morphing Guided by Perceptually Motivated Features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(8):1666–1675, Aug. 2013.

[9] P. A. Cariani and B. Delgutte. Neural correlates of the pitch of complex tones. I. Pitch and pitch salience. *Journal of neurophysiology*, 76(3):1698–1716, Sept. 1996.

[10] D. Chazan and R. Hoory. Feature-domain concatenative speech synthesis, Apr. 2006.

[11] G. Coleman and J. Bonada. Sound transformation by descriptor using an analytic domain. In *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08), Espoo, Finland, September 1-4, 2008*, page 7, 2008.

[12] G. Coleman, E. Maestre, J. Bonada, E. Maestre, and J. Bonada. Augmenting sound mosaicing with descriptor-driven transformation. In *Proceedings of DAFx-10*, page 4, 2010.

[13] N. Collins and B. L. Sturm. Sound cross-synthesis and morphing using dictionary-based methods. In *International Computer Music Conference*, 2011.

[14] G. M. Davis, S. G. Mallat, and Z. Zhang. Adaptive time-frequency decompositions. *Optical Engineering*, 33(7):2183–2191, 1994.

[15] J. Driedger and M. Müller. A Review of Time-Scale Modification of Music Signals. *Applied Sciences*, 6(2):57, Feb. 2016.

[16] J. Driedger and T. Pratzlich. Let It Bee – Towards NMF-Inspired Audio Mosaicing. In *Proceedings of ISMIR*, page 7, Malaga, 2015.

[17] H. Dudley. Thirty Years of Vocoder Research. *The Journal of the Acoustical Society of America*, 36(5):1021–1021, May 1964.

[18] A. Elowsson and A. Friberg. Long-term average spectrum in popular music and its relation to the level of the percussion. In *Audio Engineering Society Convention 142*, page 13. Audio Engineering Society, 2017.

[19] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi. Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders. In *PMLR*, July 2017.

[20] M. Goodwin. Matching pursuit with damped sinusoids. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 2037–2040, Munich, Germany, 1997. IEEE.

[21] E. Grinstein, N. Duong, A. Ozerov, and P. Perez. Audio style transfer. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 586–590, Oct. 2017.

[22] M. Hoffman and P. R. Cook. Feature-Based Synthesis: A Tool for Evaluating, Designing, and Interacting with Music IR Systems. In *Proceedings of ISMIR*, page 2, 2006.

[23] M. D. Hoffman, P. R. Cook, and D. M. Blei. Bayesian spectral matching: Turning Young MC into MC Hammer via MCMC sampling. In *ICMC*, 2009.

[24] K. Jaganathan, Y. C. Eldar, and B. Hassibi. Phase Retrieval: An Overview of Recent Developments. *arXiv:1510.07713 [cs, math]*, Oct. 2015.

[25] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60(2):115–142, July 1992.

[26] J. C. R. Licklider. A duplex theory of pitch perception. *Experientia*, 7(4):128–134, Apr. 1951.

[27] P. Mermelstein and C. Chen. Distance measures for speech recognition: Psychological and instrumental. In *Pattern Recognition and Artificial Intelligence,*, volume 101, pages 374–388. Academic Press, 1976.

[28] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1, Nov. 1993.

[29] R. Prony. Essai éxperimental et analytique: Sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, à différentes températures. *Journal de l'École Polytechnique Floréal et Plairial*, 2, 1795.

[30] L. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(1):24–33, Feb. 1977.

[31] C. Roads. *Microsound*. The MIT Press, Cambridge, Mass., Aug. 2004.

[32] D. Schwarz. State of the art in sound texture synthesis. In *Proceedings of DAFx-11*, pages 221–231, 2011.

[33] X. Serra and J. Smith. Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, 14(4):12–24, 1990.

[34] Y. Shechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase Retrieval with Application to Optical Imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, May 2015.

[35] I. Simon, S. Basu, D. Salesin, and M. Agrawala. Audio analogies: Creating new music from an existing performance by concatenative synthesis. In *Proceedings of the 2005 International Computer Music Conference*, pages 65–72, 2005.

[36] M. Slaney, M. Covell, and B. Lassiter. Automatic Audio Morphing. In *Proceedings of the Acoustics, Speech, and Signal Processing, 1996. On Conference Proceedings., 1996 IEEE International Conference - Volume 02*, volume 2 of *ICASSP '96*, pages 1001–1004, Washington, DC, USA, 1996. IEEE Computer Society.

[37] M. Slaney and R. F. Lyon. A perceptual pitch detector. In *Proceedings of ICASSP*, pages 357–360 vol.1, Apr. 1990.

[38] M. Slaney, D. Naar, and R. Lyon. Auditory model inversion for sound separation. In *Proceedings of ICASSP '94.*, volume ii, pages II/77–II/80, Adelaide, SA, Australia, 1994. IEEE.

[39] J. O. Smith. Digital audio resampling home page. Technical report, Center for Computer Research in Music and Acoustics (CCRMA), Stanford University, Jan. 2018.

[40] M. Sondhi. New methods of pitch extraction. *IEEE Transactions on Audio and Electroacoustics*, 16(2):262–266, June 1968.

[41] B. L. Sturm, C. Roads, A. McLeran, and J. J. Shynk. Analysis, visualization, and transformation of audio signals using dictionary-based methods. *Journal of New Music Research*, 38(4):325–341, Dec. 2009.

[42] W. Verhelst and M. Roelands. An Overlap-add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-scale Modification of Speech. In *Proceedings of ICASSP*, ICASSP'93, pages 554–557, Washington, DC, USA, 1993. IEEE Computer Society.

[43] P. Verma and J. O. Smith. Neural style transfer for audio spectograms. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Jan. 2018.

[44] N. Wiener. Generalized harmonic analysis. *Acta Mathematica*, 55:117–258, 1930.

[45] A. Zils and F. Pachet. Musical mosaicing. In *Proceedings of DAFx-01*, volume 2, page 135, Limerick, Ireland, 2001.