SpectroFX: Real-Time Audio Effects via Graphical Spectrogram Manipulation

Manuel Rocha[1], Pedro Duarte Pestana[2,3,4]
[1]Postgraduate at Universidade Aberta, Lisboa, Portugal, mrochapt@gmail.com
[2]Departamento de Ciências e Tecnologia, Universidade Aberta, Lisboa, Portugal
[3]Centro de Investigação em Artes e Comunicação (CIAC), Palácio de Ceia, Rua da Escola Politécnica 147, Lisboa, 1269-001, Portugal
[4]CEAUL - Centro de Estatística e Aplicações, Faculdade de Ciências, Universidade de Lisboa, Lisboa

Abstract

SpectroFX turns the short-time Fourier transform (STFT) spectrogram into a band-limited canvas for real-time audio effects inside VCV Rack. The digital signal processing (DSP) pipeline is deterministic: STFT/inverse short-time Fourier transform (ISTFT) with a periodic √Hann window and 50 % overlap ($FFT\ size\ N = 1024, hop\ H = \frac{N}{2}$), explicit $\frac{1}{N}$ IFFT scaling, and overlap-add (OLA). Magnitude operators use the Open-Source Computer Vision (OpenCV) library over a $1 \times K$ vector per frame—Gaussian blur, unsharp-mask sharpening, Sobel edge emphasis, emboss, mirror, gating and frequency stretch—and act only within a full-width horizontal band set in the widget. Phase is reconstructed by a dedicated engine with three non-iterative modes: RAW, phase vocoder (PV) and identity phase-locking (PV-Lock). Output conditioning applies a first-order high-pass direct current (DC) blocker and a soft limiter. The design privileges predictable cost and glitch-free operation.

Keywords: Spectral processing; Phase vocoder; Phase locking; VCV Rack; Real-time audio.

Título: SpectroFX: Efeitos sonoros em tempo real por manipulação gráfica

Resumo: O SpectroFX transforma o espectrograma da transformada de Fourier de tempo curto (STFT) numa "tela" limitada por banda para efeitos em tempo real no VCV Rack. O caminho de processamento digital de sinal (DSP) é determinista: STFT/transformada inversa de Fourier de tempo curto (ISTFT) com janela $\sqrt{Hann}$ e sobreposição de 50 % $\left(N = 1024, H = \frac{N}{2}\right)$, normalização explícita $\frac{1}{N}$ na IFFT e overlap-add (OLA). Os operadores de magnitude usam a biblioteca OpenCV (Open Source Computer Vision) sobre um vetor $1 \times K$ por *frame — Gaussian blur*, *unsharp mask*, realce de arestas (Sobel), *emboss*, *mirror*, *gate* e *stretch* — e atuam apenas dentro de uma banda horizontal definida no *widget*. A fase é reconstruída por três modos não iterativos: RAW, phase vocoder (PV) e *identity phase-locking* (PV-Lock) num motor dedicado. O *output* é condicionado com filtro corta-DC (alto-passa de 1.ª ordem) e *soft limiter*. A arquitetura privilegia custo previsível e operação sem perturbações,

Palavras-chave: Processamento espectral; Vocoder de fase; Bloqueio de fase; VCV Rack; Audio em tempo real.

1. Introduction

A Short Time Fourier Transform (STFT) computes the local frequency and phase content of a signal by transforming small segments, or frames, instead of the whole. Each frame captures a local slice of time and frequency, enabling "image-like" manipulations in the magnitude domain while the phase is synthesized to maintain temporal continuity. SpectroFX operationalizes this concept for performance: it limits all image processing to a full-width horizontal band selected by two draggable lines in the user interface (UI) and executes all transforms in the audio callback, ensuring that interaction stays bound by strict real-time constraints.

We intentionally target a fixed spectral geometry—FFT size ($N = 1024$), hop $H = 512$ at common sampling rates 44.1–48 kHz—because predictable timing trumps unconstrained flexibility in modular contexts. With a periodic √Hann window and 50% overlap, the constant overlap-add (COLA) condition holds, eliminating frame-boundary amplitude modulation. Using Fastest Fourier Transform in the West (FFTW) plans, we apply explicit $\frac{1}{N}$ scaling at synthesis time to maintain unity round-trip gain.

SpectroFX rejects iterative phase retrieval in the hot path. Instead, a compact phase engine offers three non-iterative options—RAW, phase vocoder (PV) and PV-Lock—covering the most relevant use-cases at bounded cost. Magnitude processing uses efficient 1D operators over the $1 \times K$ vector: Gaussian blur, unsharp mask, Sobel edge emphasis, emboss, mirror, gating and frequency stretch. Finally, a first-order high-pass DC blocker and a soft limiter stabilize levels and tame resonant peaks.

Contribution. SpectroFX contributes (i) a carefully engineered STFT/ISTFT scaffold with explicit normalization and OLA safety; (ii) a band-limited, single-frame operator stack that avoids time-coupling and guarantees bounded cost; (iii) a non-iterative phase engine combining PV and identity phase-locking; and (iv) a pragmatic set of usage recipes validated by listening tests.

Scope. We deliberately exclude free form 2D masks and iterative phase retrieval from the real-time path, and we avoid GPU requirements to keep deployment simple on modest laptops.

2. Related Work

The Short-Time Fourier Transform (STFT) remains a cornerstone technique for time-frequency analysis in audio signal processing. Classical works rigorously characterize reconstruction conditions, including the role of analysis windowing, overlap ratios, and explicit normalization to ensure Constant Overlap-Add (COLA) behavior and energy preservation, which are essential for high-fidelity spectral effects pipelines.

Foundational contributions include Allen and Rabiner's unified analysis–synthesis framework (Allen & Rabiner, 1977), Oppenheim and Schafer's systematization of discrete-time signal processing principles (Oppenheim & Schafer, 2010), and Steiglitz's practical treatment of DSP methods relevant to audio transformation (Steiglitz, 1996).

Iterative phase retrieval methods such as the Griffin–Lim algorithm reconstruct phase from magnitude-only STFTs by repeated projections (Griffin & Lim, 1984). These methods remain influential for offline workflows. However, their dependence on multiple forward–inverse transforms create non-deterministic latency and considerable computational overhead, making them unsuitable for real-time systems where bounded execution time is critical.

Phase-vocoder (PV) synthesis provides a deterministic alternative, enforcing smooth phase trajectories across frames at bounded cost. The technique has been refined extensively since its classical description in spectral audio processing (Smith, 2007), including harmonic-structure preservation strategies such as the improved time-scale modification method proposed by Laroche and Dolson (Laroche & Dolson, 1999). These enhancements improve spectral coherence and reduce transient smearing typical of early PV implementations.

With the STFT viewed as a structured two-dimensional representation, several works draw directly from image-processing concepts for spectral manipulation. Examples include smoothing, sharpening, and directional filtering approaches common in visual domains. Practical frameworks and documentation from Bradski (2000) and the OpenCV project (OpenCV, 2024) enable efficient filter-bank implementation. Similarly, spectral-domain sound-design practices and analogies to image transformations have been highlighted through creative toolkits for spectrogram-based manipulation (Charles, 2008). Research into sound texture synthesis through spectrogram modeling further reinforces the potential of cross-domain techniques (Caracalla & Roebel, 2020). Although many of these works rely on full 2D operations with high computational demand, frequency-only operators (1×K kernels) represent a relatively underexplored opportunity to retain expressiveness while satisfying the strict timing guarantees needed for live performance.

3. System Overview

SpectroFX comprises four cooperating components:

- SpectroFXModule: Manages audio I/O, circular buffers, STFT/ISTFT, operator stack, and output conditioning. Exposes processed magnitudes to the spectrogram.
- PhaseEngine: Maintains per-channel/per-bin phase state and implements RAW, PV, and PV-Lock synthesis phases using stable unwrapping and accumulation.
- Mask2D: Double-buffered structure with a lock-free front/back swap and atomic band limits lowBin/highBin; in the current DSP path, only these band limits are used (full-width horizontal mask).
- SpectroFXWidget: Code-drawn panel (no SVG) with a live spectrogram and two draggable lines that map unambiguously to integer bin indices.

Processing per hop follows: window → FFT → magnitude/phase split → 1×K magnitude operators (band-limited) → phase synthesis (RAW/PV/PV-Lock) → IFFT → window → OLA → DC-block + limiter. The UI thread never blocks the audio thread; information flows via atomics or single-writer/single-reader swaps.

Data flow detail. The input ring buffer accumulates samples until N are available for analysis; the output ring buffer releases H samples per hop, ensuring a steady cadence. Parameter snapshots (operator mixes and band indices) are read atomically once per hop to guarantee self-consistency within the frame.

4. Signal Model and STFT Formulation

Let $x[n]$ be a real-valued, discrete-time input signal. Frames are indexed by m and positioned every H samples. With an analysis window $w[n]$ of length N and $K = \frac{N}{2} + 1$ unique bins for real signals, the analysis STFT is:

$$X_m[k] = \sum_{n=0}^{N-1} x[n + mH]w[n]^{-j\frac{2\pi}{N}kn}, \qquad 0 \le k \le K-1 \tag{1}$$

STFT Analysis (per frame m, bin k)

We write $X_m[k] = A_m[k]\, e^{j\phi_m[k]}$, with magnitude $A_m[k] \ge 0$ and wrapped phase $\phi_m[k] \in [-\pi, \pi]$). For ISTFT we construct a Hermitian-symmetric spectrum $\widehat{X_m}[k]$ and compute the time-domain segment

$$\widehat{s_m}[n] = \frac{1}{N}\, w[n] \sum_{k=0}^{N-1} \widehat{X_m}[k]\ e^{j\frac{2\pi}{N}kn}, \qquad 0 \le n \le N-1 \tag{2}$$

ISTFT synthesis segment

followed by overlap-add (OLA)

$$\hat{x}[n] = \sum_{m} \widehat{s_m}[n - mH] \tag{3}$$

Overlap-add (OLA) Reconstruction

Because FFTW's IFFT omits the $1/N$ normalization, we include it explicitly to guarantee predictable levels. The bin-to-frequency map is:

$$f_k = \frac{k\, F_s}{N}, \quad 0 \le k \le K-1 \tag{4}$$

Bin-to-frequency map

and the expected phase advance between consecutive frames for a bin-center sinusoid is:

$$\Delta\phi_k^{\text{exp}} = \omega_k H = 2\pi \frac{kH}{N} \tag{5}$$

Expected inter-frame phase advance

4.1 Parseval-type energy considerations

Let $\|\cdot\|_2$ denote the (ℓ_2) norm. With a perfect reconstruction STFT pair and COLA windowing,

$$\sum_n |x[n]|^2 \approx \frac{1}{N} \sum_m \sum_{k=0}^{N-1} |X_m[k]|^2, \tag{6}$$

Parseval-type energy relation (approx.)

up to window overlap constants. Using a √Hann window ensures that squared windows tile to a constant (Section 5), simplifying gain calibration.

4.2 Window choice and leakage

We adopted the periodic Hann $h[n] = \frac{1}{2}(1 - \cos(2\pi n/N))$ and its square-root $w[n] = \sqrt{h[n]}$. The √Hann has gentle sidelobes and, crucially, pairs with itself at 50 % overlap to satisfy COLA. Leakage is inevitable for off-bin sinusoids; later we show how PV accumulation mitigates residual beating by tracking instantaneous frequency.

4.3 Spectral moments and perceptual correlations

For a magnitude vector $A[k]$, the spectral centroid and spread are:

$$C = \frac{\sum_k f_k A[k]}{\sum_k A[k]} \tag{7}$$

Spectral centroid

$$V = \sqrt{\frac{\sum_k (\boldsymbol{fk} - \boldsymbol{C})^2 \boldsymbol{A[k]}}{\sum_k A(k)}} \tag{8}$$

Spectral spread

Blur tends to lower V (smoothing), while unsharp can increase local contrast without dramatically shifting C. Although SpectroFX does not compute these metrics internally, they guide parameterization and listening tests.

4.4 Zero-padding and interpolation

We do not use zero-padding in the current real-time build. Conceptually, zero-padding by a factor Z would give $N' = ZN$ and finer bin spacing F_s/N', but at the cost of larger

FFTs and greater latency. Our design opts for fixed $N = 1024$ and relies on PV for smooth inter-frame phase evolution.

5. Windowing, Constant Overlap-Add (COLA), and Normalization

Define the periodic Hann $h[n]$ and $w[n] = \sqrt{h[n]}$. For hop $H = N/2$, the shifted windows satisfy:

$$\sum_m w[n - mH]^2 \; = \; \sum_m h[n - mH] \; = \; \text{const}, \quad \forall n \tag{9}$$

COLA sum (√Hann, 50% overlap)

A brief derivation: the periodic Hann can be written as a sum of complex exponentials,

$$h[n] = \frac{1}{2} - \frac{1}{4} e^{j\frac{2\pi}{N}n} - \frac{1}{4} e^{-j\frac{2\pi}{N}n} \tag{10}$$

Hann as complex-exponential sum

Shifting by $H = \frac{N}{2}$ flips the sign of the exponentials, and summing two shifts recovers a constant. Since $w^2 = h$, the COLA sum is constant as well. Consequently, OLA reconstructs a flat envelope without breathing at frame boundaries.

Because the IFFT is unnormalized in many libraries, we multiply by $1/N$ on synthesis. We validate normalization via two canonical tests:

a) Impulse test $x[n] = \delta[n]$: after STFT/ISTFT, OLA yields a train of √Hann-shaped segments whose sum equals the unit impulse (modulo the high-pass and limiter if engaged).

b) Bin-center sine $x[n] = \sin(2\pi f_k n/F_s)$: reconstruction should exhibit unity gain and no beating; any residual modulation indicates window or scaling errors.

We also note that window symmetry and index centering matter in practice; our implementation uses consistent buffer indexing so that analysis and synthesis windows align.

6. Phase Reconstruction: RAW, PV, and PV-Lock

Let $\phi_m[k]$ be the analysis phase and $\widehat{\phi_m}[k]$ the synthesis phase. The three modes are:

6.1 RAW (analysis phase reuse)

A direct copy:

$$\widehat{\phi_m^{\text{RAW}}}[k] \; = \; \phi_m[k] \tag{11}$$

RAW mode: reuse analysis phase

RAW preserves transient crispness when magnitude edits are light, but under strong edits on sustained tones it can sound diffuse because inter-frame phase continuity is not enforced.

6.2 Phase vocoder (PV) (inter-frame continuity)
We estimate instantaneous frequency per bin via phase deviation from the expected advance. Let $princarg(\theta) \in [-\pi, \pi]$ denote the principal argument. The unwrap step computes:

$$\Delta\phi_m^{\text{dev}}[k] = princarg\left((\phi_m[k] - \phi_{m-1}[k]) - \Delta\phi_k^{\text{exp}}\right) \tag{12}$$

Phase deviation (principal-value unwrap)

then the instantaneous radian frequency is

$$\widehat{\omega_m}[k] \;=\; \frac{\Delta\phi_k^{\text{exp}} + \Delta\phi_m^{\text{dev}}[k]}{H} \tag{13}$$

Instantaneous angular frequency (per bin)

The synthesis phase accumulates:

$$\widehat{\phi_m^{\text{PV}}}[k] \;=\; \widehat{\phi_{m-1}}[k] + \widehat{\omega_m}[k]\,H, \qquad \text{with initial condition } \phi_0^{\text{PV}}[k] = \phi_0[k] \tag{14}$$

Phase-vocoder accumulation

This enforces smooth inter-frame evolution even when magnitudes change.
Numerical note. We keep phases wrapped at each step to avoid drifting and store them in single precision per bin; accumulation error remains negligible at audio rates with typical session lengths.

6.3 PV-Lock (identity phase-locking) (harmonic coherence)
PV can slightly chorus harmonic stacks because each bin evolves independently. Identity phase-locking improves coherence by anchoring non-peak bins to the nearest spectral peak. Let $\mathcal{P}_m$ be the peak set at frame m (local maxima above a small relative threshold). For a given bin k, define $p(k) \in \mathcal{P}_m$ as the nearest peak (tie-break to the larger magnitude). The locked phase propagates local PV differences outward from peaks:

$$\widehat{\phi_m^{\text{LOCK}}}[p] \;=\; \widehat{\phi_m^{\text{PV}}}[p], \quad p \in \mathcal{P}_m \tag{15}$$

Identity phase-locking at spectral peaks

and for neighbors (e.g., to the right),

$$\widehat{\phi_m^{\text{LOCK}}}[k] \;=\; \widehat{\phi_m^{\text{LOCK}}}[k-1] \;+ princarg\left(\widehat{\phi_m^{\text{PV}}}[k] - \widehat{\phi_m^{\text{PV}}}[k-1]\right) \tag{16}$$

Identity phase-locking propagation to neighbors

In practice we apply the same rule to the left side separately or simply integrate differences radially from each peak within a limited neighborhood. The complexity is $O(K)$ per frame.

6.4 Synthesis spectrum

Given edited magnitudes $\widehat{A_m}[k]$ (Section 7), the synthesis spectrum is:

$$\widehat{X_m}[k] \;=\; \widehat{A_m}[k]\; e^{j\,\widehat{\phi_m}[k]} \tag{17}$$

Synthesis spectrum from magnitude and phase

We enforce $\widehat{A_m}[k] \geq \varepsilon$ for a small floor $\varepsilon \sim 10^{-6}$ to avoid pathological zeros which can destabilize angle unwrapping between frames.

Aside: time-scale modification. The classic PV supports time-stretch by a factor α via a modified hop $H_s = \alpha H$ and appropriate phase advance correction. SpectroFX uses $\alpha = 1$ (no time-scaling) to keep latency and cost fixed, but the derivation clarifies why the same machinery yields stable synthesis.

7. Magnitude-Domain Operators (1×K) with OpenCV

At each hop, we hold the magnitude vector $a \in R^{K}_{\geq 0}$ and a band $\mathcal{B} = \{k : k_l \leq k \leq k_h\}$ defined by the UI lines. Each operator maps $a \mapsto t$. We blend inside the band with $\alpha \in [0,1]$:

$$\hat{A}[k] \;=\; \begin{cases} (1-\alpha)\, a[k] + \alpha t[k], & k \in \mathcal{B}, \\ a[k], & otherwise \end{cases} \tag{18}$$

In-band blend rule (magnitude operator mix)

7.1 Gaussian blur (spectral smoothing)

A discrete Gaussian kernel g_σ produces:

$$t[k] \;=\; (g_\sigma * a)[k] = \sum_{r=-R}^{R} g_\sigma[r]\; a[k-r], \qquad g_\sigma[r] \propto \exp\left(-\frac{r^2}{2\sigma^2}\right) \tag{19}$$

Gaussian blur along frequency

We implement with OpenCV's separable Gaussian using a σ mapped from the Blur control. Blur reduces spectral roughness and narrow spikes, often yielding smoother *highs* (cymbals, hiss).

7.2 Unsharp mask (detail enhancement)

Subtract a blurred version and add a scaled residue:

$$t[k] \;=\; a[k] \;+\; \beta\,(a[k] - (g_\sigma * a)[k]), \quad \beta \geq 0 \tag{20}$$

Unsharp mask

This boosts local contrast without significantly shifting energy. Excessive β may excite the limiter; modest values are recommended.

7.3 Edge emphasis (Sobel-type derivative)
Using a centered difference,

$$d[k] \;=\; a[k+1] - a[k-1], \qquad t[k] = |d[k]| \tag{21}$$

Edge emphasis (centered difference)

Edges emphasize rapid spectral changes (formant boundaries, consonant bursts). We normalize by a small constant to keep scales comparable across material.

7.4 Emboss (tilted difference kernel)
A simple 3-tap kernel, as an example:

$$h[-1] = -2, \qquad h[0] = -1, \qquad h[1] = 1,$$

produces:

$$t[k] = \sum_{r=-1}^{1} h[r]\; a[k-r] \tag{22}$$

Emboss (three-tap tilt)

followed by affine rescaling to $[0, \infty]$. Sonically this adds a tilted metallic sheen that can enliven textures.

7.5 Mirror (Nyquist symmetry reflection)
Reflect around Nyquist:

$$k^* = (K-1) - k, \qquad t[k] = a[k^*] \tag{23}$$

Mirror around Nyquist

Within $\mathcal{B}$, mirror generates complementary timbral colors.

7.6 Gate (relative thresholding)
With $a_{max} = \; max_k \, a\,[k]$ and $\theta = \gamma a_{max} (\gamma \in [0,1])$

$$t[k] = \begin{cases} (1-\delta)\, a[k], & a[k] \;<\; \theta, \\ a[k], & otherwise, \end{cases} \qquad \delta \in [0,1] \tag{24}$$

Gate with relative threshold

Removes weak/noisy bands; useful on sustained pads or noisy recordings.

7.7 Stretch (frequency resampling)

A factor $s > 0$ rescales bin positions:

$$t[k] = Interp(a,\, s^{-1}k) \tag{25}$$

Stretch via frequency-axis resampling

OpenCV's resize handles interpolation. $s > 1$ expands spacing (brighter, airier), $s < 1$ compresses (darker, denser). Because phase is synthesized per bin, moderate stretch factors remain coherent under PV/PV-Lock.

7.8 Safeguards, ordering, and saturation

Operators are stateless across frames and commute approximately. Using an order like blur → unsharp/edge/emboss → gate → mirror/stretch (or the opposite order, when necessary). After blending, we clamp $\hat{A}[k] \leftarrow \max\left(\hat{A}[k], \varepsilon\right)$ to avoid zeros. A final soft limiter in time-domain contains narrow peaks.

Parameter mapping. GUI controls are mapped linearly to $\sigma, \beta, \gamma, \delta, s$ within musically useful ranges. For instance, $\sigma \in [0.5,5]$ bins, $\beta \in [0,1]$, $s \in [0.8,1.25]$. Exact ranges can be tuned per preset without altering complexity.

8. Band Mask and UI Mapping

The UI exposes two vertical lines mapped to bin indices k_l, k_h. The mapping from screen space to frequency is linear in bins but non-linear in Hz:

$$k_l = \left\lfloor \frac{N}{F_s} f_l \right\rfloor, \qquad k_h = \left\lfloor \frac{N}{F_s} f_h \right\rfloor, \qquad 0 \le k_l \le k_h \le K - 1 \tag{26}$$

UI band limits mapped to bin indices

We store *lowBin*, *highBin* atomically; the audio thread snapshots them once per hop. A history buffer used by the spectrogram (HIST ≈ 256 columns) is maintained on the UI side and never read by DSP, guaranteeing that drawing cannot stall audio.

Practical note. Users often prefer logarithmic frequency tick marks; the internal mapping remains integer bins, which makes the band selection unambiguous and stable across zoom levels.

Ergonomics. The two-line model also prevents accidental "holes" or zig-zag shapes that free painting would create, improving predictability on stage.

9. Latency, Real-Time Constraints, and Complexity

Latency. With centered windows, algorithmic delay is approximately $N/2$ samples; an additional initial offset (on the order of N) ensures OLA continuity from the first hop.

Host buffering adds to this, so end-to-end latency is:

$$\tau_{\text{total}} \approx \frac{N}{2F_s} + \tau_{\text{host}} \tag{27}$$

Approximate end-to-end latency

Complexity. Per hop and per channel:
c) FFT/IFFT: $O(N \log N)$.
d) Operators: $O(K)$ per operator with small constants.
e) PV/PV-Lock: $O(K)$ (peak detection plus phase propagation).

Because $\mathcal{B}$ spans a full width and we avoid 2D convolution, work is independent of spectrogram history length. Memory is dominated by a handful of N-sized buffers and per-bin phase/magnitude arrays; the footprint is modest.

Denormal handling. We avoid denormal slowdowns by ensuring the high-pass filter and limiter keep samples away from sub-normal ranges, and by recommending CPU (central processing unit) flush-to-zero if available.

Scheduling. Audio callbacks are single-threaded and time-bounded; all allocations are done during construction. OpenCV routines used are non-allocating on steady state (operating on pre-allocated Mats).

10. Implementation Details (Buffers, Plans, and Safety)

FFTW plans are created once per channel using FFTW_MEASURE; pointers are reused across hops to eliminate allocation churn.

√Hann window vectors are precomputed (double precision) and applied multiplicatively at analysis and synthesis.

Circular buffers decouple per-sample I/O from per-hop block operations. Read/write indices wrap modularly and are advanced by H.

Hermitian completion. We store K bins and mirror the rest at synthesis to form an N-point complex spectrum with conjugate symmetry.

Precision. Magnitudes are stored as float; phases as float wrapped via “atan2 / arg” logic. Internal accumulators can be double if sessions are extremely long, though single precision has proved sufficient.

Threading. The audio thread owns all DSP state. UI changes are communicated through atomics or a lock-free swap. No OpenCV threading is invoked in the audio thread.

Parameterization and control voltage. All mixes map knobs/control voltage (CV) to [0,1]. The phase mode is an enumerated selector (0=RAW, 1=PV, 2=PV-Lock), avoiding ambiguous mid-values.

Safety margins. We keep ≈ −6 dB nominal headroom before the limiter to accommodate sharp kernels (unsharp/edge).

Error handling. If FFT plan creation fails or buffers misalign, the module falls back to a bypass path (implementation detail) to guarantee continuity.

Numerical stability. Angle wrapping uses princarg to avoid cumulative drift. A small magnitude floor ε prevents undefined angles. Window multiplication occurs before FFT and after IFFT to preserve COLA.

11. Output Conditioning: DC-Blocker and Soft Limiter

A first-order high-pass DC blocker removes slow bias drift:

$$y[n] \;=\; x[n] - x[n-1] + R\,y[n-1], \qquad 0 < R < 1 \tag{28}$$

First-order DC blocker

For $F_s \in [44.1{,}48]$ kHz and $R \approx 0.995$, the −3 dB cutoff lies near 35–38 Hz:

$$f_c \;\approx\; \frac{(1-R)\,F_s}{2\pi} \tag{29}$$

Approximate cutoff frequency of the DC blocker

After OLA, the signal is passed through a soft limiter,

$$\tilde{y}[n] \;=\; \tanh(g\,y[n]) \tag{30}$$

Soft-limiter nonlinearity

with drive g tuned for gentle containment rather than heavy saturation. The limiter is not a mastering tool; it prevents additive peaks from clipping when kernels increase local contrast. Users should still gain-stage appropriately.

Perceptual note. The limiter's odd-symmetric transfer largely preserves waveform shape for small g, minimizing coloration. For clarity-critical material, set g low and rely on upstream gain staging.

12. Evaluation: Objective Checks and Listening Notes

We advocate a minimal but effective test suite:

1. Impulse response. Verify OLA smoothness and lack of discontinuities at hop boundaries. With just the high-pass and limiter on mild settings, the reconstructed impulse train should sum close to the original.

2. Bin-center sine. At $f_k = kF_s/N$, the round-trip is unity gain; with PV/PV-Lock, transitions across frames are smooth. RAW also works here but can highlight leakage if magnitude edits are aggressive.

3. Off-bin sine and sweep. A logarithmic sweep checks leakage behavior and confirms absence of spurious lines; PV accumulation minimizes beating artifacts.

4. Mode toggle on sustained tones. PV-Lock usually yields the most "solid" harmonics; PV is slightly airier; RAW retains a raw, transient-friendly quality.

5. Operator sweeps. Explore parameter ranges for blur (σ), unsharp (β), edge/emboss gains, gate (γ, δ), and stretch s. Note interactions with the limiter.

Listening notes. For percussive loops, RAW and PV often sound similar unless edits are extreme; RAW can feel punchier. For tonal music, PV-Lock most reliably preserves harmonic structure; PV can gently diffuse. For noisy pads, operator choice dominates: blur smooths, unsharp adds presence, mirror introduces novel colors.

Objective proxies. While formal perceptual metrics (e.g., PESQ) are out of scope, simple proxies - log-spectral distance (LSD) before/after, spectral flatness change - can be computed offline to characterize operator effects on reference material.

13. Use Cases and Presets

Three reproducible presets illustrate typical workflows and parameter regimes.

(A) Air polish (4–12 kHz).
$k_l = \lfloor 4\ \text{kHz} \cdot N/F_s \rfloor), (k_h = \lfloor 12\ \text{kHz} \cdot N/F_s \rfloor$. Gaussian blur $\sigma \approx 2$ bins, unsharp $\beta \approx 0.3$, blend $\alpha \approx 0.25$. Phase = PV-Lock. Outcome: reduced harshness with perceived openness.
Rationale. Cymbals and hiss exhibit high variance; blur reduces roughness while unsharp restores definition of broader features.

(B) Body enhancer (150–600 Hz).
Moderate unsharp ($\beta \approx 0.4$) and light stretch $s \in [0.95, 1.05]$. Phase = PV. Outcome: firmer fundamentals and controlled bloom in bass/mid-bass.
Rationale. Low-mid regions benefit from detail emphasis; a tiny stretch subtly tilts spectral spacing to avoid masking.

(C) Texture edge (1–6 kHz).
Edge emphasis blended at $\alpha \approx 0.4$, optional emboss for character. Phase = RAW for drums; PV-Lock for sustained textures. Outcome: definition in noisy or layered material.
Rationale. Edge operators accentuate rapid spectral transitions (consonants, transients) while PV-Lock preserves overall harmonic continuity for sustained material.

Workflow tips. Start with broad bands; make small moves; monitor limiter activity; compare phase modes on looped sections; commit presets with descriptive names that encode band and operator choices.

14. Discussion and Conclusions

SpectroFX demonstrates that real-time spectral manipulation can be achieved with a deterministic architecture that preserves both computational predictability and perceptual coherence. By leveraging a COLA-safe STFT/ISTFT scaffold with explicitly normalized overlap-add synthesis, the system ensures stable gain and glitch-free operation at fixed latency. The non-iterative phase engine enables three controllable behaviors: RAW for transient-rich material, the classical phase vocoder (PV) for smooth spectral continuity, and identity phase-locking (PV-Lock) for improved harmonic stability. These modes constitute a pragmatic design space that supports live sound design without the latency penalties associated with iterative methods.

A key trade-off of the approach is the deliberate exclusion of free-form 2D masks within the real-time DSP path. Restricting operations to horizontal band-defined, frame-local 1×K kernels guarantees constant cost and simplifies user interaction, although it limits the range of spatially complex effects available. Transient smearing is modest but non-zero when strong magnitude edits are combined with PV; a hybrid onset-aware strategy could further preserve percussive clarity. Peak detection stability directly affects PV-Lock behavior, especially on dense or noise-like spectra, which motivates future work on locally adaptive or hysteresis-based peak selection.

The current implementation avoids GPU dependencies and maintains CPU cache locality, which simplifies deployment on modest hardware. Future extensions may explore GPU-accelerated convolution or deferred processing for non-real-time modes, provided that determinism in the live signal path is preserved. Additional frequency-only operators such as median smoothing or adaptive equalization remain lightweight opportunities for expansion.

Overall, SpectroFX contributes a reproducible and extensible method for live graphical spectrogram manipulation. The design produces musically useful transformations across a variety of source materials while maintaining bounded execution time within a modern modular-synthesis context. The system establishes a robust foundation for further user-driven evolution, particularly through validation studies assessing both perceived audio quality and workflow benefits in performance scenarios.

Code availability: The SpectroFX source code is open-source and available at the GitHub repository: **https://github.com/mrochapt/SpectroFX**

Acknowledgments: We thank the VCV Rack community and early users for testing and feedback. This work was supported by FCT - Fundação para a Ciência e a Tecnologia, Portugal, under projects UID/04019/2025.

REFERENCES

Allen, J. B., & Rabiner, L. R. (1977). A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE, 65*(11), 1558–1564. https://doi.org/10.1109/PROC.1977.10770

Audio Engineering Society. (n.d.). *AES E-Library*. https://www.aes.org/e-lib/

Bradski, G. (2000). The OpenCV library. *Dr. Dobb's Journal of Software Tools*.

Caracalla, H., & Roebel, A. (2020, May). Sound texture synthesis using RI spectrograms. In *ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 416–420). IEEE.

Charles, J. F. (2008). A tutorial on spectral sound processing using Max/MSP and Jitter. *Computer Music Journal, 32*(3), 87–102.

FFTW. (2024). *Fastest Fourier Transform in the West*. http://www.fftw.org/

Franzoni, V. (2023). Cross-domain synergy: Leveraging image processing techniques for enhanced sound classification through spectrogram analysis using CNNs. *Journal of Autonomous Intelligence, 6*(3), 1–14.

Frigo, M., & Johnson, S. G. (2005). The design and implementation of FFTW3. *Proceedings of the IEEE, 93*(2), 216–231.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*(2), 236–243. https://doi.org/10.1109/TASSP.1984.1164317

Laroche, J., & Dolson, M. (1999). Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing, 7*(3), 323–332.

Liu, X., Delany, S. J., & McKeever, S. (2019). Sound transformation: Applying image neural style transfer networks to audio spectrograms. In *Computer Analysis of Images and Patterns (CAIP 2019)* (pp. 330–341). Springer.

Lukin, A., & Todd, J. (2006, May). Adaptive time-frequency resolution for analysis and processing of audio. In *Proceedings of the 120th Audio Engineering Society Convention*. Audio Engineering Society.

OpenCV. (2024). *OpenCV documentation*. https://docs.opencv.org/

Oppenheim, A. V., & Schafer, R. W. (2010). *Discrete-time signal processing* (3rd ed.). Prentice Hall.

Smith, J. O. (2007). *Spectral audio signal processing*. W3K Publishing. https://ccrma.stanford.edu/~jos/sasp/

Steiglitz, K. (1996). *A digital signal processing primer: With applications to digital audio and computer music*. Addison-Wesley.

VCV. (2025). *VCV Rack plugin API guide*. https://vcvrack.com/manual/PluginGuide
Zölzer, U. (Ed.). (2011). *DAFX: Digital audio effects* (2nd ed.). Wiley.

Manuel Rocha (ORCID: 0009-0008-2785-7668) holds a degree in Computer Engineering from Universidade Aberta, where he is also pursuing advanced studies in Web Technology. He maintains active interests in digital signal processing and real-time computational systems, and is the author of *SpectroFX*, a VCV Rack plugin for real-time spectrogram manipulation. With more than twenty years of experience in enterprise technology, he is Head of IT at a major multinational company, where he leads the strategic direction, governance and transformation of large-scale information systems. His career includes the management of mission-critical infrastructures, the coordination of high-performance engineering teams and the implementation of complex integration and modernization initiatives across mobility, transport and corporate domains. He has extensive experience in real-time systems, cloud and datacenter operations, cybersecurity and the design of resilient, future-proof technology architectures.

Pedro Duarte Pestana (ORCID: 0000-0002-3406-1077) holds a PhD in Music Informatics from the Catholic University of Portugal. He serves as an Assistant Professor in the Section of Informatics, Physics and Technology (SIFT) at Universidade Aberta since 1 March 2022. His work lies at the intersection of multimedia signal processing and the perception and cognition of auditory and visual phenomena. He has worked as a consultant for several international signal-processing software companies and contributed to the formation of a Montreal-based startup focused on artificial-intelligence systems for audio processing. He publishes regularly in venues such as the Audio Engineering Society, DAFx and ISMIR, for which he is also an active reviewer and member. He is Portugal's national representative for ISO/TC 43/SC 1. He was the recipient of a Gulbenkian Professorship between 2015 and 2019 and served as Director of the Research Centre for Science and Technology of the Arts (CITAR), where he is currently a collaborating researcher. He is also an integrated researcher at the Research Centre for Arts and Communication – Universidade Aberta (CIAC-UAb). CiênciaVitae: https://www.cienciavitae.pt/portal/2714-8A7B-5CCA