SOUND MATCHING WITH A DIFFERENTIABLE KARPLUS-STRONG ALGORITHM

Pablo Tablas de Paula[*] *, David Marttila, Rodrigo Díaz,*
Irán Román, Emmanouil Benetos and Joshua D. Reiss

Centre For Digital Music
Queen Mary University Of London
London, UK
p.tablasdepaula@qmul.ac.uk

ABSTRACT

We present a self-supervised, event-based sound matching model using a differentiable extended Karplus-Strong algorithm. To avoid relying on external onset and fundamental frequency detectors, we explore training methodologies combining parameter losses on synthetic data with audio losses. We demonstrate that time-domain fractional delay interpolation provides gradient accuracy comparable to frequency-sampling while avoiding time-aliasing in highly resonant time-varying scenarios. Through systematic gradient analysis, we reveal that standard spectral losses provide no meaningful directional gradients for onset times, heavily degrading joint training. Training exclusively with parameter losses on synthetic data effectively learns fundamental frequency, timbral parameters, and onset times, but struggles to generalise to monophonic studio recordings of plucked guitar. External detectors combined with audio losses generalise best, isolating the model to timbre optimisation. While our Karplus-Strong decoder recovers interpretable parameters and naturally captures the transient characteristics of plucked guitar, Harmonics plus Noise baselines yield higher reconstruction fidelity by most metrics.

1. INTRODUCTION

The Karplus-Strong Algorithm (KSA) offers an efficient method to simulate plucked instruments [1]. Originally conceived as a self-changing wavetable, Jaffe and Smith re-interpreted it as a fedback delay line with additional filtering to simulate dispersion [2], an insight that would inspire the wider field of Digital Waveguide (DWG) Synthesis and enable physical modelling in real-time [3].

Engel et al. demonstrated that DSP parameters can be optimised end-to-end with neural networks via Differentiable Digital Signal Processing (DDSP) [4], a self-supervised analysis-by-synthesis paradigm in which controls are learned directly from audio, sparking growing interest in extending the technique beyond the original Harmonics plus Noise (HpN) decoder [5, 6]. The KSA is a natural candidate: it requires far fewer parameters than HpN and naturally captures transient characteristics. In the online supplement to Hayes et al., audio losses were used to optimise the decay of a frequency-sampling Linear Time Invariant (LTI) implementation of KSA [7]. More recently, we extended this to a fully time-domain implementation in which decay, timbre, pluck-position, and dynamics are free to vary at every frame [8].

However, both studies relied on external f_0 and onset detectors; the first are prone to octave errors due to pitch fluctuations around transients, while the latter introduce time-shifts, missed onsets and confounds string-fret interactions with plucks. Additionally, this frame-based parameter prediction is unnecessary for discrete plucks: assuming expressive techniques such as vibrato or pitch bends are absent, a pluck's damping, decay, and f_0 are constant over its note.

Optimising these parameters by gradient descent requires fully differentiable architectures. In the frequency domain, fractional delays become continuous phase shifts [9, 10, 11], but time-aliasing is severe when impulse responses exceed the FFT frame size [12]. Time-domain interpolation avoids this and is now computationally viable through parallelised automatic differentiation [13]; however, its gradient behaviour under spectral losses is unexplored.

Such losses are notoriously unreliable for pitch [14] and, as we show, give no useful gradients for onset times. Recent DDSP work stabilises frequency prediction with Parameter Loss ($\mathcal{P}_{loss}$), the distance between parameters of synthetic audio made by the synthesiser in question [15, 16]; we extend this to onset supervision. Our broader aim is a self-supervised event-based model: one that recovers the discrete structure of a performance (how many plucks, when, and at what pitch) without hand-annotated onsets or f_0, jointly with the string timbre, deriving its supervision instead from the differentiable synthesiser and freely-labelled synthetic data.

Our contributions are threefold:

- We show that delay line length is differentiable in the time domain via Lagrange Interpolation, with gradient behaviour comparable to frequency sampling and without time-aliasing artefacts.
- We perform a systematic gradient analysis showing that standard spectral losses fail to provide meaningful directional gradients for onset times, heavily degrading joint audio-parameter optimisation.
- We demonstrate that $\mathcal{P}_{loss}$ on synthetic data effectively learns f_0, timbre, and onset times, while external detectors combined with audio losses generalise best to real-world data.

2. BACKGROUND

2.1. Karplus-Strong Synthesis

The KSA circulates a short broadband noise burst, typically one period of the desired Fundamental Frequency (f_0), through a dig-

[*] Supported by the UKRI Centre for Doctoral Training in Artificial Intelligence and Music (grant EP/S022694/1.)

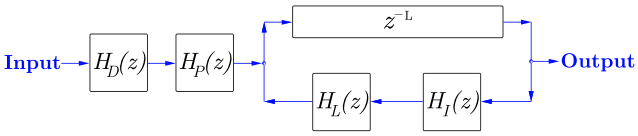

Figure 1: An extended Karplus-Strong Algorithm.

ital delay line in a feedback loop, the burst and delay line being analogous to pluck and string. The loop delay L (in possibly fractional samples) is the round-trip time of a transversal wave along the string, setting the perceived f_0 via

$$L = \frac{f_s}{f_0}, \tag{1}$$

where f_s is the sampling frequency; this assumes delay-free filters, with the loop filter's phase delay later compensated in D. In the feedback path, a *loop filter* $H_L(z)$ models frequency-dependent string losses. The original wavetable used a two-point average $H_L(z) = \frac{1}{2}(1 + z^{-1})$; we adopt a more modern approach proposed by Välimäki et al., which favours the expressiveness of Infinite Impulse Response (IIR) filters:

$$H_L(z) = g\frac{1 + a_1}{1 + a_1 z^{-1}}. \tag{2}$$

Here $H_L(z)$ is implemented as a first-order all-pole filter and re-parameterised to provide separate, decoupled control over the overall decay time via the DC gain g, and frequency-dependent damping via the cutoff coefficient a_1 [17].

Because the ideal delay length L rarely resolves to an integer number of samples, an interpolation filter $H_I(z)$ in the feedback path makes the delay fractional. All-pass filters avoid the strong low-pass effect of linear interpolation [2], but their recursive nature makes them prone to transients when the delay changes. We instead use *Lagrange Interpolation*, whose fractional delay is far more uniform across frequency, with a maximally flat magnitude response in the lower spectrum [18]. The coefficients $h[n]$ for the Lagrange interpolator are computed as:

$$h[n] = \prod_{\substack{k=0 \\ k\neq n}}^{N} \frac{D-k}{n-k}, \quad n = 0, 1, \ldots, N, \tag{3}$$

where N is the order of the Finite Impulse Response (FIR) filter and D the desired delay. We use $N = 5$, which keeps the fractional delay accurate while limiting passband distortion [18]. The final Lagrange interpolator has the form:

$$H_I(z) = \sum_{n=0}^{N} h[n] z^{-n}. \tag{4}$$

Since $H_L(z)$ adds its own frequency-dependent phase delay, the fractional delay D fed to the Lagrange interpolator must absorb both the non-integer remainder of L and the loop-filter phase delay $P_L(f_0)$ [2]:

$$D = L - \lfloor L \rfloor - P_L(f_0), \quad \text{where} \quad P_L(f_0) = -\frac{\angle H_L\left(e^{j\omega_0}\right)}{\omega_0}. \tag{5}$$

Before entering the delay line, the noise burst is shaped by two excitation filters [2]. The *pluck position filter* models where along the string it is struck: plucking at a fractional length imposes nodes that suppress certain harmonics, captured by a feedforward comb $H_P(z) = 1 - z^{-p}$ with p proportional to the pluck position. Being feedforward and applied once, its fractional delay reduces to linear interpolation:

$$H_P(z) = 1 - z^{-p_i}\left((1 - p_f) + p_f\, z^{-1}\right), \tag{6}$$

where $p_i = \lfloor p \rfloor$ and $p_f = p - \lfloor p \rfloor$.

Lastly, since stronger impacts introduce more high-frequency energy, Jaffe and Smith model intensity with a *dynamics filter* [2], which acts perceptually as a strike intensity rather than a distance control:

$$H_D(z) = \frac{1 - R}{1 - Rz^{-1}}, \tag{7}$$

where $R \in [0, 1]$ is computed from the fundamental frequency f_0 and the desired dynamic level to keep perceived loudness uniform across the range (derivation in [2]). The complete extended algorithm is shown in Figure 1.

2.2. Synthesiser Sound Matching

Sound matching is the inverse of synthesis: given a target sound, recover the parameters that reproduce it. For physical models like the KSA this is also analytical; the predicted parameters give interpretable insight into the object's physical properties [19]. Early, costly evolutionary methods [19] gave way to supervised Neural Network (NN) regression [20] minimising $\mathcal{P}_{loss}$; however, $\mathcal{P}_{loss}$ optimises numerical rather than acoustic accuracy, often yielding inconsistent perceptual reconstructions [16].

2.3. Differentiable Digital Signal Processing

DDSP implements signal processors as differentiable operations, turning the synthesiser into an interpretable decoder so an audio-reconstruction loss can backpropagate to its controls and train end-to-end [4]. Its standard objective, Multi-Scale Spectral Loss ($\mathcal{L}_{MSS}$), measures L_1 distance between amplitude spectrograms, effective for timbre but unreliable for pitch [14], often needing external f_0 predictors [4]. Recent work mitigates this via Wasserstein-based Spectral Optimal Transport ($\mathcal{L}_{SOT}$) for frequency shifts [21], revised $\mathcal{L}_{MSS}$ configurations [22], and $\mathcal{P}_{loss}$ pre-training [16].

2.4. Gradient-Based Optimisation for Recursive Filters

Naive time-domain implementation of recursive filters requires sample-by-sample calculation, generating prohibitively large computational graphs. Two complementary approaches restore parallelism.

The *frequency sampling method* evaluates the filter response over M linearly-spaced frequencies:

$$z_M = \left[e^{j\pi\frac{0}{M-1}}, e^{j\pi\frac{1}{M-1}}, \ldots, e^{j\pi}\right] \tag{8}$$

yielding a parallelisable FIR approximation via DFT ratio of feedforward and feedback coefficients. Arbitrarily complex filter structures (series connections and recursive loops) are composed through elementwise products and matrix inversions over z_M [12], and fractional delays D (in samples) reduce to a smooth and differentiable closed-form phase shift [9]:

$$z_M^{-D} = e^{-j2\pi k M^{-1} D}. \tag{9}$$

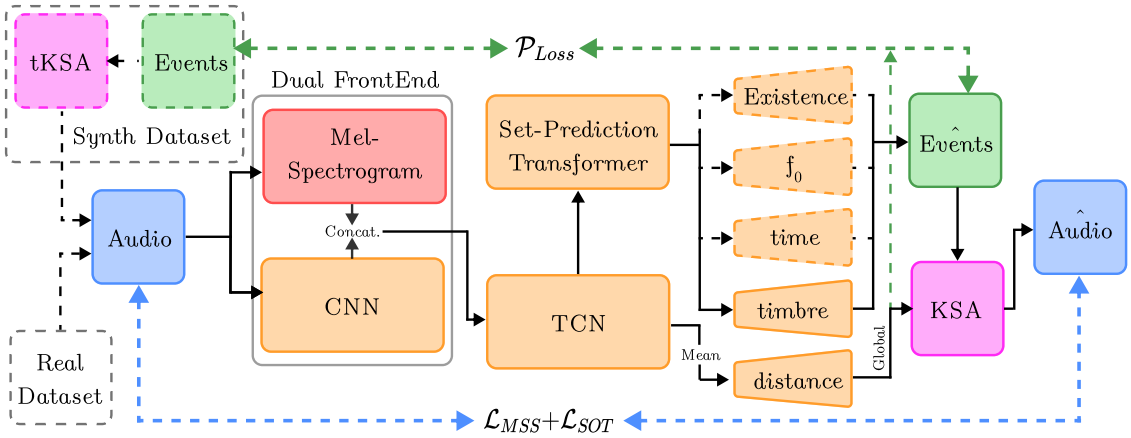

Figure 2: Proposed sound matching pipeline. Dashed lines indicate optional paths defining our three training regimes: $\mathcal{P}$-Only (end-to-end prediction trained with $\mathcal{P}_{loss}$ uniquely on Synth Data), $\mathcal{P}$+Audio (end-to-end prediction trained with both $\mathcal{P}_{loss}$ and audio losses), and Audio-Only (trained solely with audio losses, bypassing the existence, f_0 and onset time parameter heads with external detectors).

A Hermitian symmetry constraint enforces a real-valued impulse response of length $N_{\text{FFT}} = 2(M-1)$; if the true impulse response exceeds this, circular convolution introduces time-aliasing. For time-varying parameters, the DDSP literature combines frequency sampling with frame-based Overlap-add (OLA) [4, 6, 11].

Alternatively, Yu and Fazekas restore native time-domain parallelism by computing exact analytical gradients for recursive filters via the adjoint method, eliminating both memory bottlenecks and the truncation artefacts inherent to frequency sampling [13]. Fractional delays in this setting have been handled by backpropagating through discrete interpolators [23], but to the knowledge of the authors, not through audio losses.

2.5. Set-Based Prediction

We base our event encoder on the Sound Event Detection Transformer (SEDT) [24], an event-based, end-to-end model that frames detection as direct set prediction: a fixed set of learnable queries is matched to the ground-truth events in a single pass, regressing each event's boundaries without frame-wise post-processing or autoregressive decoding. We adapt this paradigm to jointly estimate discrete pluck events and their physical parameters.

3. DIFFERENTIABLE KARPLUS-STRONG

Because the KSA is a recursive, energy-dissipating system, predicting parameters on a continuous frame-by-frame basis can cause errors to accumulate through the feedback loop. Provided no extended techniques are present in the data, it is more robust to use an event-based model, representing acoustic variations as discrete note-level events with constant damping, decay, and f_0. Figure 2 illustrates the complete pipeline, consisting of an Event-Based Encoder (§3.1) and two differentiable KSA decoder variants (§3.2): a Time-Domain (tKSA) and a Frequency-Sampling (fKSA) decoder.

3.1. Event-Based Encoder

The encoder maps a monophonic waveform to a fixed-size set of N=10 event parameter slots through four stages. We set N=10 as headroom: most samples contain only one or two plucks, but a few NSynth recordings (e.g. tempo-synced ones; [25]) contain more.

Dual-Frontend Feature Extractor. Inspired by [26], the audio (4 s, f_s=16 kHz; §4.2) is processed in parallel by a log-magnitude Mel filterbank (1024-point FFT, 256-sample hop, 128 bands, 32–8000 Hz) for f_0 and harmonic structure and a learnable 1D CNN (4 layers, up to 64 channels, stride 4) for phase-sensitive transients; each is followed by batch norm and ReLU, then truncated to a common time axis and concatenated.

Temporal Convolutional Network. The concatenated features are projected to $d_{\text{model}} = 64$ and processed by a 7-block Temporal Convolutional Network (TCN) with exponentially increasing dilations $(1, 2, \ldots, 64)$, building a global contextual memory representation.

Set-Prediction Transformer. $N = 10$ learnable query embeddings attend to the TCN memory via 3 standard Transformer decoder layers (8-head self- and cross-attention, GELU feed-forward), with sinusoidal positional encoding, layer normalisation, and dropout ($p = 0.1$).

Parameter Projection Heads. Each query embedding is decoded by shared linear heads predicting: event existence (sigmoid logit); onset time (sigmoid-mapped scalar); f_0 (soft-argmax over 128 log-spaced bins, 32–2000 Hz); and four timbral parameters: decay (g), damping (a_1), pluck position (p), and dynamic level, each sigmoid-mapped to physical bounds. A separate head predicts a global distance gain from the mean TCN representation.

3.2. KSA Decoders

At inference, and for both $\mathcal{P}_{loss}$ training and synthetic dataset generation, we use a buffer-based, sample-by-sample, non-differentiable implementation of the KSA. Training under audio losses uses a differentiable onset by placing a noise burst via a Straight Through Estimator (STE): the forward pass snaps to the nearest integer sample, while the backward pass routes gradients through the continuous phase rotation of Eq. 9, giving smooth directional gradients for onset time. This excitation method is paired with the following differentiable KSA implementations:

Time-Domain (tKSA). Using parallelised backpropagation through all-pole filters [27], the loop filter $H_L(z)$, Lagrange interpolator $H_I(z)$, and integer delay $z^{-L_{int}}$ are combined into a

single closed-loop transfer function:

$$H_{tKSA}(z) = \frac{H_D(z)H_P(z)}{1 - H_L(z)H_I(z)z^{-L_{int}}}, \tag{10}$$

where $L_{int} = \lfloor L \rfloor$.

Frequency Sampling (fKSA). Discrete interpolators and integer delays are replaced by the continuous fractional phase shift of Eq. 9:

$$H_{fKSA}(z_M) = \frac{H_D(z_M)H_P(z_M)}{1 - H_L(z_M)z_M^{-L}}. \tag{11}$$

Owing to the highly resonant nature of KSA (Figure 3), no single N_{FFT} avoids time-aliasing for all configurations while preserving time-varying dynamics. We find $N_{FFT} = 2^{14}$ with a 256-sample Hann-windowed hop an effective compromise for the decay and onset timings of our dataset (predominantly a sustained note at 0s with a palm mute at 3s; see appendix A for sound examples), though time-aliasing still occurs (Fig. 4).

4. TRAINING

4.1. Losses

The total training objective combines two audio losses with a parameter loss, each defined in the subsections below:

$$\mathcal{L} = w_{\text{MSS}}\,\mathcal{L}_{\text{MSS}} + w_{\text{SOT}}\,\mathcal{L}_{\text{SOT}} + w_{\mathcal{P}}\,\mathcal{P}_{\text{Loss}}. \tag{12}$$

4.1.1. Audio Losses

We adopt the *Revisited* $\mathcal{L}_{MSS}$ of Schwär and Müller [22], a *vertical* (point-wise) loss [21] computing the ℓ_2-distance between log-compressed magnitude spectrograms at six scales (prime window lengths $\{67, 127, 257, 509, 1021, 2053\}$, flat-top window, no zero-padding, γ=1):

$$\mathcal{L}_{\text{MSS}} = \sum_{s=1}^{6} \big\| \log(1+\gamma|S_s|) - \log(1+\gamma|\hat{S}_s|) \big\|_2. \tag{13}$$

$\mathcal{L}_{SOT}$ complements this with a *horizontal* comparison [21]: within each Short-Time Fourier Transform (STFT) frame it measures the squared 2-Wasserstein cost of transporting magnitude energy *along* the frequency axis, providing directional gradients for frequency shifts where the vertical $\mathcal{L}_{MSS}$ fails [14]:

$$\begin{aligned} \mathcal{L}_{\text{SOT}} &= \sum_t W_2^2\Big(\tilde{S}_t,\, \tilde{\hat{S}}_t\Big)\,, \\ W_2^2(p,q) &= \int_0^1 \big(F_p^{-1}(u) - F_q^{-1}(u)\big)^2\, du, \end{aligned} \tag{14}$$

where $\tilde{S}_t$ is the magnitude spectrum of frame t normalised into a probability distribution over frequency and F^{-1} is its quantile (inverse-CDF) function. We adopt the best-performing configuration and weights of Torres et al. (w_{MSS}=0.05, w_{SOT}=1.0) [21], but pair them with the smooth Revisited $\mathcal{L}_{MSS}$ above.

4.1.2. Parameter Loss ($\mathcal{P}_{loss}$)

When ground-truth parameters are available (synthetic data), we optimise a set-prediction loss, following SEDT [24]. At each step, Hungarian matching assigns the N predicted slots to $M \leq N$ ground-truth events using a cost matrix combining temporal L_1 and existence probability; unmatched slots receive zero-padded no-event targets. The matched loss is:

$$\begin{aligned} \mathcal{P}_{\text{Loss}} = w_e\,\mathcal{L}_e + w_t\,\mathcal{L}_t + w_{f_0}\,\mathcal{L}_{f_0} + w_g\,\mathcal{L}_g \\ + w_{\text{rt60}}\,\mathcal{L}_{\text{rt60}} + w_p\,\mathcal{L}_p + w_d\,\mathcal{L}_d \end{aligned} \tag{15}$$

where the weights $(w_e, w_t, wf_0, w_g, w_{\text{rt60}}, w_p, w_d)$ are set to $(100, 10, 2, 1, 1, 0.5, 0.5)$, respectively. Existence ($\mathcal{L}_e$) uses sigmoid focal loss [28] with standard hyperparameters (α=0.5, γ=2) to handle class imbalance. Onset time, pluck position, dynamic level, and global distance gain ($\mathcal{L}_t, \mathcal{L}_p, \mathcal{L}_d, \mathcal{L}_g$) use L_1 losses on sigmoid-activated predictions. Fundamental frequency ($\mathcal{L}_{f_0}$) applies negative log-likelihood over K=128 log-spaced bins.

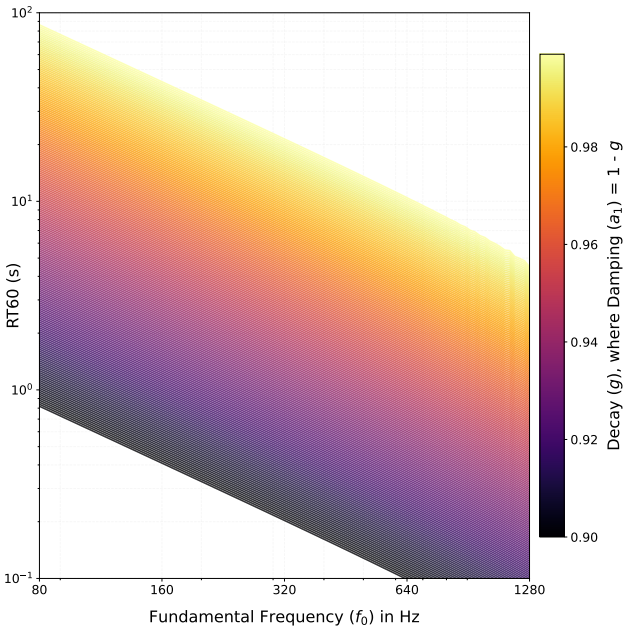

Figure 3: RT60 values across f_0, damping (a_1), and decay (g) in tKSA (without $H_P(z)$ and $H_D(z)$), illustrating their highly coupled nature. At the most resonant of the computed combinations ($g = 0.999$ and $a_1 = 0.001$), the RT60 is 86.26s at $f_0 = 80$Hz, while being only 4.5s at $f_0 = 1280$Hz.

For RT60 ($\mathcal{L}_{\text{rt60}}$), rather than supervising the coupled g and a_1 independently, we compute their joint analytical decay time. A signal at harmonic k circulates through $H_L(e^{j\omega_k})$ exactly f_0 times per second, giving:

$$\text{RT60}_k = \frac{-3}{f_0 \log_{10}|H_L(e^{j\omega_k})|}, \tag{16}$$

where the -3 in the numerator reflects the -60 dB (factor 10^{-3}) amplitude decay that defines RT60, evaluated for the first 10 harmonics below Nyquist. An L_1 loss in $\log(1 + \text{RT60})$ space provides balanced gradients across short and long decays.

4.2. Data

Synthetic Dataset (Synth). Samples are 4-second mono clips at f_s=16 kHz containing 1–10 pluck events at constant f_0 over D1–D6, with per-event parameters sampled from Beta distributions and a 30% palm-mute probability. Full generation details

and reproducible code can be found in appendix A. The training set comprises 8360 samples re-seeded each epoch; a fixed 290-sample validation set monitors convergence.

Real-World Dataset (Real). We use the acoustic guitar subset of NSynth [25], retaining non-reverberant samples in the D1–D6 range, yielding 4-second mono recordings at 16 kHz split as [8360 / 930 / 290] for train / validation / test.

4.3. Models

We evaluate seven models across three regimes. **$\mathcal{P}$-Only** uses the non-differentiable tKSA trained with $\mathcal{P}_{loss}$ alone, predicting all event parameters (existence, onset, f_0, timbre) end-to-end. **Audio-Only** models use differentiable KSA decoders with audio losses only, replacing the existence, onset, and f_0 heads with external detectors: CREPE [29] (Viterbi, tiny) for f_0 and Spectral Flux with backtracking [30] for onsets; the signal is segmented at detected onsets and each event's f_0 set to the median over its confidently voiced frames (CREPE confidence > 0.5). **$\mathcal{P}$+Audio** models use the same differentiable decoders trained end-to-end with both $\mathcal{P}_{loss}$ and audio losses ($\mathcal{L}_{MSS}$, $\mathcal{L}_{SOT}$), jointly predicting all parameters; we further ablate them by detaching the audio-loss gradient from the onset, f_0, or both heads. All KSA variants are evaluated in both frequency-sampling (freq.) and time-domain (time) forms. **HpN** and **HpN+** are Harmonics plus Noise baselines using the original DDSP encoder [4] and our TCN encoder respectively, both trained with audio losses and external detectors (CREPE for f_0, A-weighted loudness for dynamics), without reverberation.

4.4. Training Process

All models use AdamW (lr=10^{-3}, weight decay 10^{-4}, batch size 32) with gradient clipping at norm 1.0 and ReduceLROnPlateau (factor 0.5, patience 15, min lr=10^{-6}), on A100 GPUs[1] with early stopping (patience 50, threshold 10^{-4}, max 24 hrs or 10000 epochs). $\mathcal{P}$+Audio training adds the parameter loss at $w_{\mathcal{P}}$=1.0 to the audio-loss weights of §4.1 (w_{MSS}=0.05, w_{SOT}=1.0), which Audio-Only models use alone. When mixing datasets, epoch sizes are matched and dataloaders combined via max-size cycling.

5. EVALUATION

We evaluate on two 290-sample test sets: held-out synthetic and NSynth acoustic guitar. For Audio-Only models on synthetic data, detectors run on-the-fly to mirror training. Metrics fall into three categories.

Signal Fidelity. To assess signal fidelity, we employ $\mathcal{L}_{MSS}$ and $\mathcal{L}_{SOT}$ identically to their usage during training (Eqs. 13, 12). We also evaluate perceptual timbral distance using a warped Mel-Frequency Cepstral Coefficients (MFCC) metric, which calculates the Dynamic Time Warping (DTW)-normalised L_1 distance between 20-coefficient MFCC sequences; the dynamic time warping naturally absorbs minor temporal misalignments. Additionally, we measure the cosine similarity between the target and predicted Root Mean Square (RMS) energy envelopes to evaluate the overall envelope fidelity.

Parameter and Event Reconstruction. Because ground-truth parameters are required, these metrics are computed exclusively on the synthetic test set. We evaluate event detection performance using onset precision, recall, and F1 scores by matching predicted slots that have an existence probability greater than 0.5 against ground-truth onsets within a 25ms window. Pitch accuracy is measured by calculating the MAE f_0 in cents (Δ Cents) across all matched active events. Finally, we report the overall parameter loss ($\mathcal{P}_{\text{Loss}}$) as previously defined in Eq. 15.

Distributional and Perceptual Similarity. For the real-world NSynth dataset, we compute Kernel Audio Distance (KAD) [31], a distribution-free Maximum Mean Discrepancy metric that, unlike the Fréchet Audio Distance, remains strictly unbiased even at small sample sizes. Following the methodology of [32], we calculate KAD across two distinct embedding spaces to capture different perceptual qualities: EnCodec [33] is used to measure low-level spectro-temporal fidelity, while CLAP [34] provides an assessment of higher-level semantic and timbral alignment.

Table 1: Single-event gradient accuracy (%) comparing $\mathcal{L}_{MSS}$ and $\mathcal{L}_{SOT}$ across tKSA and fKSA.

	$\mathcal{L}_{MSS}$				$\mathcal{L}_{SOT}$			
	tKSA		fKSA		tKSA		fKSA	
Parameter	CGA	FGA	CGA	FGA	CGA	FGA	CGA	FGA
f_0	68.0	58.4	69.6	63.6	54.8	50.8	50.0	50.0
g (decay)	100.0	100.0	100.0	99.6	70.4	47.6	73.6	47.6
a_1 (damping)	100.0	100.0	99.2	94.8	40.0	50.0	34.4	50.0
Pluck position	66.8	89.2	67.6	90.0	49.6	48.0	47.6	46.0
Dynamic level	100.0	99.6	98.4	92.4	97.6	100.0	98.0	98.8
Distance gain	100.0	100.0	96.4	82.0	99.6	99.2	97.2	86.8
Onset time	51.2	55.2	52.0	56.0	42.4	14.4	62.4	88.8

6. EXPERIMENTS

Gradient Analysis. We evaluate whether gradients point toward the target parameter:

$$\operatorname{sgn}\left(\frac{\partial \mathcal{L}(\hat{x}f[n], x[n])}{\partial f}\right) = \operatorname{sgn}(f - f_x) \tag{17}$$

Table 1 reports Coarse Gradient Accuracy (CGA, 250 uniform initialisations across the full parameter range) and Fine Gradient Accuracy (FGA, 250 initialisations within $\pm 10\%$ of target) for a single event at $\{f_0, t, a_1, g, p, \text{gain}, \text{dyn}\} = \{110\,\text{Hz}, 2\,\text{s}, 0.2, 0.99, 0.25, 0.9, 0.9\}$, across both losses ($\mathcal{L}_{\text{MSS}}$, $\mathcal{L}_{\text{SOT}}$) and implementations (tKSA, fKSA). Table 2 extends this to multi-event sequences ($K \in [2, 5]$) focusing on the onset time parameter. To maintain a consistent evaluation budget of approximately 250 initialisations regardless of sequence length, we construct a uniform K-dimensional grid where the number of test points per dimension scales as $\lfloor 250^{1/K} \rceil$ (e.g., $16^2 = 256$ for $K = 2$, $6^3 = 216$ for $K = 3$). We report *Any* accuracy (gradient points toward any ground-truth event) and *Joint* accuracy (all K predictions simultaneously pulled toward their Hungarian-matched targets).

Synthetic Parameter Recovery. All five KSA models (§4.3) are trained and evaluated on synthetic data, scored with signal-fidelity and parameter-reconstruction metrics (§5); results are reported in Table 3.

Real-World Instrument Modelling. $\mathcal{P}$+Audio models train on mixed Synth+Real data, Audio-Only and the HpN baselines on Real data; $\mathcal{P}$-Only is unchanged from the previous experiment. Evaluation uses the NSynth acoustic guitar test set with signal-fidelity and distributional metrics (§5); results are reported in Table 4.

[1]This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045

Table 2: Multi-event onset time gradient accuracy (%) for K equally spaced plucks. fKSA uses $N_{FFT} = 16384$, hop=256. *Any*: at least one gradient points toward a target; *Joint*: all K simultaneously toward their Hungarian-matched targets.

	$\mathcal{L}_{MSS}$				$\mathcal{L}_{SOT}$			
	tKSA		fKSA		tKSA		fKSA	
Events	Any	Joint	Any	Joint	Any	Joint	Any	Joint
2	73.4	22.7	77.1	23.0	64.1	12.1	77.5	28.5
3	83.3	7.4	82.7	13.9	79.9	5.6	84.9	11.6
4	71.6	1.6	72.8	4.7	70.7	1.2	77.6	6.6
5	65.9	0.4	65.3	3.3	66.9	1.2	68.6	4.1

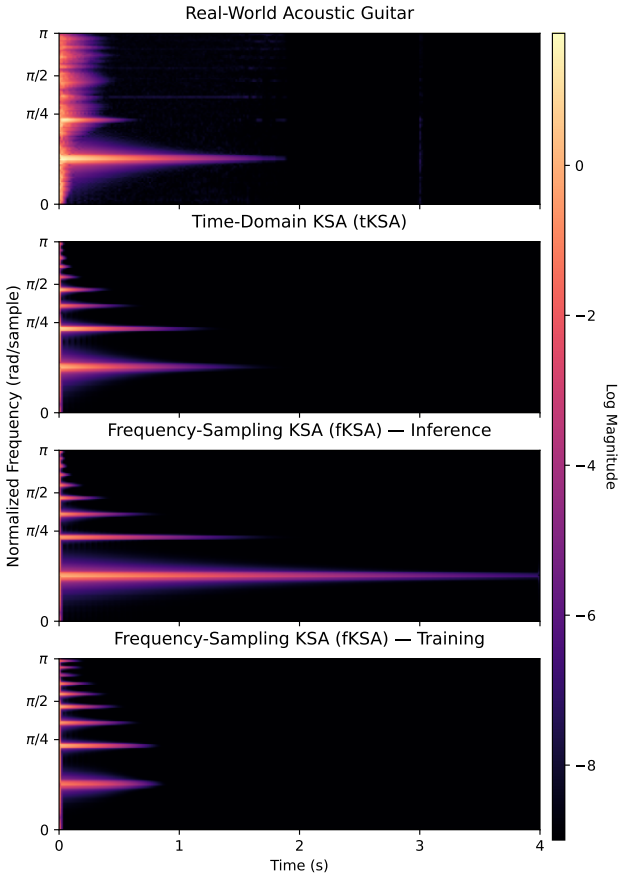

Figure 4: Sound matching on an NSynth acoustic guitar recording. Top: target. Middle: tKSA and fKSA Audio-Only reconstructions at inference. Bottom: the same fKSA reconstruction rendered via the frequency-sampling forward pass used in training, where time-aliasing from circular convolution at $N_{FFT} = 2^{14}$ truncates long decays, leading the model to overestimate decay g and underestimate damping a_1, visible as excess sustain above.

7. DISCUSSION

Table 1 confirms that time-domain interpolation gradients are comparable to frequency-sampling, validating tKSA as a viable alternative. As a *vertical*, point-wise distance, $\mathcal{L}_{MSS}$ provides reliable gradients for timbral and dynamics parameters but struggles with pluck position. The *horizontal* $\mathcal{L}_{SOT}$, which only redistributes energy along the frequency axis, is weaker on most parameters and, surprisingly given its design, blind to f_0; we hypothesise this is due to the fact that KSA delay length influences the decay, as shown in Figure 3. Its one striking exception is onset time under fKSA, where it reaches the highest single-event accuracy of any configuration (88.8% FGA), against near-blindness on the alias-free tKSA (14.4%). A plausible explanation is that frequency-sampling time-aliasing (Fig. 4) couples a temporal shift into the per-frame spectral content that the horizontal $\mathcal{L}_{SOT}$ can resolve, an artefact rather than a robust onset cue; however further experiments would be needed to confirm this.

Critically, this single-event advantage does not survive to realistic sequences. Table 2 reports *Joint* onset accuracy, all K onsets pulled toward their matched targets at once, collapsing for every loss as K grows (to at most 4.1% at K=5), as the losses distribute events unevenly and yield unusable temporal structure. No audio loss, then, provides usable gradients for the onsets of a multi-event signal. This explains the consistent underperformance of $\mathcal{P}$+Audio ablations relative to $\mathcal{P}$-Only and Audio-Only counterparts across both experiments, evidenced by the low RMS values[2]. This is corroborated by the $\mathcal{P}$+Audio ablations, where detaching the onset head from audio losses always results in strongly increased performance. We additionally hypothesise that time-aliasing is the primary factor further degrading frequency-sampling relative to time-domain implementations in Audio-Only variants, where identical onset and f_0 targets isolate the aliasing effect. As shown in Fig. 4, decays whose impulse responses exceed the FFT frame length are truncated by circular convolution, causing the fKSA model to underestimate sustain.

The $\mathcal{P}$-Only ablation performs best on synthetic data, even for timbre. In addition to being given a clear direction towards the global optima, we attribute this to the perceptual RT60 reparameterisation (see Eq. 16), which jointly supervises damping and decay. Δ Cents error is inflated by octave misalignments in Audio-Only models, which also show lower onset recall.

However, $\mathcal{P}_{loss}$ generalises poorly to real-world data, frequently missing onsets, likely due to CNN overfitting to synthetic transient features and erratically predicting f_0. Future work could narrow this gap by augmenting the synthetic data with a realistic noise floor and phase distortion from randomised all-pass filters to emulate microphone capture, by supervising f_0 and onset times with a higher-fidelity physical simulation such as a finite-difference time-domain model, or, most directly, by training on annotated real-world recordings.

As expected, the HpN baselines achieve superior reconstruction across most metrics. Unsurprisingly, this gap reflects a structural ceiling of the KSA: its single feedback loop captures neither the resonant guitar body, nor the interaction between the string and the player's hands, nor string–fret collisions, all contributions the HpN model can absorb directly and that future research should address through more comprehensive physical models. Although we argue that our proposed KSA decoder represents transient characteristics more faithfully than the HpN variants, we acknowledge that increasing the frame resolution of HpN is likely to further improve its ability to reconstruct transients and that no formal listening test was conducted. Instead, we encourage readers to listen to the audio examples in the accompanying repository (appendix A).

Lastly, designing audio losses that yield useful gradients for event *location*, and for f_0, remains a key open challenge. Ver-

[2]We hypothesise the low $\mathcal{L}_{SOT}$ in these collapsed models is an artefact of per-frame low-pass filtering normalisation, resulting in unreliable metrics for reconstructions with wrongly-predicted silent frames.

Table 3: Synthetic parameter recovery. All models are trained and evaluated on synthetic data. $\mathcal{P}$-Only uses tKSA with $\mathcal{P}_{\text{Loss}}$ supervision; $\mathcal{P}$+Audio models use differentiable decoders trained end-to-end with both $\mathcal{P}_{\text{Loss}}$ and audio losses ($\mathcal{L}_{\text{MSS}}$, $\mathcal{L}_{\text{SOT}}$); Audio-Only models use differentiable decoders with audio losses and external detectors for onsets and f_0. Rows beneath each $\mathcal{P}$+Audio model are stop-gradient ablations that detach the audio-loss gradient from the named head (f_0, onset, or both), leaving it supervised only by $\mathcal{P}_{\text{Loss}}$. Per column, the best, second-best, and third-best values are shown in **bold**, underline, and *italic* respectively.

	$\mathcal{L}_{MSS}\downarrow$	$\mathcal{L}_{SOT}\downarrow$	wMFCC $\downarrow$	RMS $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$	Δ Cents $\downarrow$	$\mathcal{P}_{loss}\downarrow$
$\mathcal{P}$-Only	**136.11**	0.18	**6.02**	**0.96**	0.92	**0.92**	**0.92**	15.14	**0.29**
Audio-Only (freq.)	158.95	*0.19*	9.29	0.91	**0.94**	0.85	0.89	294.39	17.49
Audio-Only (time)	149.07	**0.16**	9.23	*0.93*	**0.94**	0.85	0.89	294.33	15.66
$\mathcal{P}$+Audio (freq.)	252.90	*0.19*	15.29	0.33	0.01	0.01	0.01	116.96	4.72
$\hookrightarrow$ detach f_0	251.79	0.22	15.09	0.36	0.02	0.02	0.02	45.49	3.73
$\hookrightarrow$ detach onset	195.29	0.27	10.92	0.81	0.31	0.31	0.31	42.60	1.94
$\hookrightarrow$ detach both	163.00	0.21	*8.15*	0.94	0.83	0.83	0.83	**12.69**	*0.52*
$\mathcal{P}$+Audio (time)	241.16	0.20	13.47	0.43	0.01	0.01	0.01	*13.86*	3.74
$\hookrightarrow$ detach f_0	245.77	0.18	16.29	0.38	0.01	0.01	0.01	216.76	4.55
$\hookrightarrow$ detach onset	157.95	0.20	8.19	*0.93*	0.64	0.64	0.64	13.15	0.82
$\hookrightarrow$ detach both	*155.96*	0.23	7.59	0.94	*0.84*	*0.84*	*0.84*	14.26	0.48

Table 4: Real-world instrument modelling on the NSynth acoustic-guitar test set. KSA regimes are as in Table 3, but with $\mathcal{P}$+Audio trained on mixed Synth+Real data and Audio-Only on Real data; HpN and HpN+ are Harmonics plus Noise baselines (original DDSP and our TCN encoder) trained on Real data. KAD is reported in CLAP and EnCodec embedding spaces. Ablations and highlighting follow Table 3.

	$\mathcal{L}_{MSS}\downarrow$	$\mathcal{L}_{SOT}\downarrow$	wMFCC $\downarrow$	RMS $\uparrow$	CLAP $\downarrow$	EnCodec $\downarrow$
HpN	246.60	*0.14*	7.90	*0.96*	266.20	17.99
HpN+	**183.43**	**0.02**	**6.20**	**1.00**	183.11	**5.20**
$\mathcal{P}$-Only	311.86	0.20	13.18	0.95	193.78	64.40
Audio-Only (freq.)	299.47	0.30	14.08	0.93	*191.50*	71.91
Audio-Only (time)	*262.72*	0.20	*12.18*	0.97	**163.95**	*40.39*
$\mathcal{P}$+Audio (freq.)	468.33	0.13	32.43	0.09	547.78	494.88
$\hookrightarrow$ detach f_0	466.40	0.20	30.80	0.12	431.52	482.48
$\hookrightarrow$ detach onset	393.13	0.28	21.44	0.86	363.81	118.33
$\hookrightarrow$ detach both	348.18	0.25	15.37	0.81	204.19	116.66
$\mathcal{P}$+Audio (time)	461.06	0.18	26.63	0.16	403.83	470.67
$\hookrightarrow$ detach f_0	463.00	0.18	27.29	0.16	446.45	480.10
$\hookrightarrow$ detach onset	336.72	0.19	14.88	0.82	213.53	96.80
$\hookrightarrow$ detach both	341.44	0.21	15.08	0.83	201.14	99.01

tical, point-wise distances cannot indicate which way spectral energy should move; optimal-transport formulations, in theory, should. Beyond $\mathcal{L}_{SOT}$ along the frequency axis [21], a *joint* time–frequency transport cost [35], growing with displacement in both time and frequency, could in principle resolve onset and f_0 *together*.

8. CONCLUSION

We presented an event-based sound-matching system built on a differentiable extended Karplus-Strong algorithm, comparing time-domain and frequency-sampling decoders under end-to-end and external-detector training. Time-domain Lagrange interpolation matches frequency-sampling gradient accuracy while avoiding frame-based time-aliasing, making it a viable, artefact-free alternative for differentiable physical modelling of time-varying highly-resonant systems.

Our gradient analysis identifies the absence of directional onset gradients from spectral losses as the primary failure mode of joint audio-parameter training. Parameter loss on synthetic data supervises f_0, timbre, and onsets well, but learned detectors do not yet generalise to real transients; external detectors with audio losses remain most robust on real data, though Harmonics-plus-Noise baselines retain higher reconstruction fidelity on most metrics. The exception is the CLAP perceptual distance, on which our time-domain Audio-Only model surpasses both baselines; we hypothesise this stems from the KSA's natural modelling of sharp transients, though formal listening tests are needed to confirm it. All audio examples are in appendix A.

Future work should close the sim-to-real onset gap via training on higher-quality synthetic and annotated acoustic data, and explore joint time-frequency optimal-transport losses, which measure spectral energy displacement along both axes simultaneously [35, 21]. Extending the decoder to a more comprehensive physical model that simulates string-fret collisions and string-player interactions is the natural path toward higher-fidelity guitar reconstruction.

A. AUDIO EXAMPLES AND CODE

Audio examples for all synthetic parameter recovery and real-world instrument modelling experiments, along with the reference code for reproducing experiments, figures, synthetic data generation, and NSynth data filtering, are available at `https://ptablasdpaula.github.io/DAFx26-Karplus/`.

B. REFERENCES

[1] K. Karplus and A. Strong, "Digital Synthesis of Plucked-String and Drum Timbres," *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983.

[2] D. A. Jaffe and J. O. Smith, "Extensions of the Karplus-Strong Plucked-String Algorithm," *Computer Music Journal*, vol. 7, no. 2, pp. 56–69, 1983.

[3] P. Tablas de Paula, J. O. Smith III, V. Välimäki, and J. D. Reiss, "Four Decades of Digital Waveguides," *J. Audio Eng. Soc.*, 2026, in press.

[4] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *Proc. of ICLR*, Addis Ababa, Ethiopia, 2020.

[5] F. Caspe, A. McPherson, and M. Sandler, "DDX7: Differentiable FM Synthesis of Musical Instrument Sounds," in *Proc. of ISMIR*, Bengaluru, India, 2022.

[6] D. Südholt, M. Cámara, Z. Xu, and J. D. Reiss, "Vocal Tract Area Estimation by Gradient Descent," in *Proc. of DAFx*, Copenhagen, Denmark, 2023.

[7] B. Hayes, J. Shier, G. Fazekas *et al.*, "A Review of Differentiable Digital Signal Processing for Music and Speech Synthesis," *Front. Signal Process.*, vol. 3, p. 1284100, 2024.

[8] P. Tablas de Paula, D. Marttila, and J. D. Reiss, "Differentiable Karplus-Strong," in *Digital Music Research Network One-Day Workshop (DMRN+20)*, London, UK, 2025.

[9] S.-C. Pei and Y.-C. Lai, "Closed Form Variable Fractional Time Delay Using FFT," *IEEE Signal Processing Letters*, vol. 19, no. 5, pp. 299–302, 2012.

[10] A. I. Mezza, R. Giampiccolo, A. Bernardini *et al.*, "Differentiable Scattering Delay Networks for Artificial Reverberation," in *Proc. of DAFx*, Ancona, Italy, 2025, pp. 202–207.

[11] A. Carson, A. Wright, and S. Bilbao, "Gradient-based Optimisation of Modulation Effects," *J. Audio Eng. Soc.*, 2026, in press.

[12] G. Dal Santo, G. M. De Bortoli, K. Prawda *et al.*, "FLAMO: An Open-Source Library for Frequency-Domain Differentiable Audio Processing," in *Proc. of IEEE ICASSP*, Hyderabad, India, 2025, pp. 1–5.

[13] C.-Y. Yu and G. Fazekas, "Accelerating Automatic Differentiation of Direct Form Digital Filters," in *1st Workshop on Differentiable Systems and Scientific Machine Learning at EurIPS*, Copenhagen, Denmark, 2025.

[14] J. Turian and M. Henry, "I'm Sorry for Your Loss: Spectrally-Based Audio Distances Are Bad at Pitch," in *NeurIPS Workshop on "I Can't Believe It's Not Better!"*, Vancouver, BC, Canada, 2020.

[15] J. Engel, R. Swavely, L. H. Hantrakul *et al.*, "Self-Supervised Pitch Detection by Inverse Audio Synthesis," in *ICML Workshop on Self-supervision in Audio and Speech*, Vienna, Austria, 2020.

[16] N. Masuda and D. Saito, "Improving Semi-supervised Differentiable Synthesizer Sound Matching for Practical Applications," *IEEE/ACM TASLP*, vol. 31, pp. 863–875, 2023.

[17] V. Välimäki, J. Huopaniemi, M. Karjalainen, and Z. Jánosy, "Physical Modeling of Plucked String Instruments with Application to Real-Time Sound Synthesis," *J. Audio Eng. Soc.*, vol. 44, no. 5, pp. 331–353, 1996.

[18] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the Unit Delay [FIR/All Pass Filters Design]," *IEEE Signal Processing Magazine*, vol. 13, no. 1, pp. 30–60, 1996.

[19] J. Riionheimo and V. Välimäki, "Parameter Estimation of a Plucked String Synthesis Model Using a Genetic Algorithm with Perceptual Fitness Calculation," *EURASIP JASP*, vol. 2003, no. 8, p. 758284, 2003.

[20] L. Gabrielli, S. Tomassetti, S. Squartini *et al.*, "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection," in *Proc. of DAFx*, Edinburgh, UK, 2017.

[21] B. Torres, G. Peeters, and G. Richard, "Unsupervised Harmonic Parameter Estimation Using Differentiable DSP and Spectral Optimal Transport," in *Proc. of IEEE ICASSP*, Seoul, South Korea, 2024, pp. 1176–1180.

[22] S. Schwär and M. Müller, "Multi-scale Spectral Loss Revisited," *IEEE Signal Processing Letters*, vol. 30, pp. 1712–1716, 2023.

[23] C. Churchwell, M. Kim, and P. Smaragdis, "Combolutional Neural Networks," in *Proc. of IEEE WASPAA*, New Paltz, NY, USA, 2025, pp. 1–5.

[24] Z. Ye, X. Wang, H. Liu *et al.*, "Sound Event Detection Transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[25] J. Engel, C. Resnick, A. Roberts *et al.*, "Neural Audio Synthesis of Musical Notes with WaveNet Autoencoders," in *Proc. of ICML*, Sydney, NSW, Australia, 2017, pp. 1068–1077.

[26] J. Shier, F. Caspe, M. Sandler *et al.*, "Differentiable Modelling of Percussive Audio with Transient and Spectral Synthesis," in *Proc. of Forum Acusticum*, Turin, Italy, 2023.

[27] C.-Y. Yu, C. Mitcheltree, A. Carson *et al.*, "Differentiable All-Pole Filters for Time-Varying Audio Systems," in *Proc. of DAFx*, Guildford, Surrey, UK, 2024, pp. 345–352.

[28] T.-Y. Lin, P. Goyal, R. Girshick *et al.*, "Focal Loss for Dense Object Detection," in *Proc. of IEEE ICCV*, Venice, Italy, 2017, pp. 2980–2988.

[29] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "CREPE: A Convolutional Representation for Pitch Estimation," in *Proc. of IEEE ICASSP*, Calgary, AB, Canada, 2018, pp. 161–165.

[30] B. McFee *et al.*, "Librosa: Audio and Music Signal Analysis in Python," in *Proc. 14th Python in Science Conf. (SciPy)*, Austin, TX, USA, 2015, pp. 18–24.

[31] Y. Chung, P. Eu, J. Lee *et al.*, "KAD: No More FAD! An Effective and Efficient Evaluation Metric for Audio Generation," in *AI Heard That! ICML Workshop on Machine Learning for Audio*, Vancouver, BC, Canada, 2025.

[32] M. Tailleur, J. Lee, M. Lagrange *et al.*, "Correlation of Fréchet Audio Distance with Human Perception of Environmental Audio Is Embedding Dependent," in *Proc. of EUSIPCO*, Lyon, France, 2024, pp. 56–60.

[33] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High Fidelity Neural Audio Compression," *Trans. Mach. Learn. Res.*, 2023.

[34] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP Learning Audio Concepts from Natural Language Supervision," in *Proc. of IEEE ICASSP*, Rhodes Island, Greece, 2023, pp. 1–5.

[35] L. Fabiani, S. J. Schlecht, and F. Elvander, "Time-Frequency Audio Similarity Using Optimal Transport," in *Proc. 58th Asilomar Conf. on Signals, Systems, and Computers*, 2024, pp. 1414–1417.