FROM MUSIC AUDIO TO EXPRESSIVE DYNAMICS CONTROL: TRANSCRIBING MIDI CC1 FROM REAL MUSIC PERFORMANCES VIA DIFFERENTIABLE ANALYSIS-BY-SYNTHESIS

Yifan XIE[1] and **Mathieu BARTHET**[2, 1]

[1]**Centre for Digital Music, Queen Mary University of London**, London, England
[2]**Aix Marseille Univ, CNRS, PRISM**, Marseille, France

ABSTRACT

We study automated transcription of the MIDI CC1 control trajectory from real bowed-string audio, where CC1 in sample-based orchestral libraries (e.g., Spitfire Audio's BBC SO) controls dynamics while also influencing timbral aspects through crossfading between dynamic layers. This setting lacks ground-truth CC1 annotations and suffers from a synthetic-to-real domain gap, making direct supervised learning on real recordings infeasible. We propose a differentiable analysis-by-synthesis framework that adapts a CC1 transcription network to unlabelled real audio via a learned surrogate renderer. The surrogate maps an aligned MIDI score and a CC1 trajectory to a compact expressive feature space ϕ consisting of root mean square (RMS) energy and spectral centroid (SC), enabling stable gradients for adaptation to real recordings without CC1 annotations. On 34 violin recordings from the URMP dataset, analysis-by-synthesis evaluation using BBC SO yields an RMS-envelope correlation of $r = 0.844$, improving over the model trained only on synthetic data ($r = 0.675$) on all 34 pieces ($p < 10^{-6}$, Cohen's $d = 1.90$). We further observe that RMS-only adaptation can harm timbral alignment ($r_{\text{CC1-SC}} = -0.087$), while including spectral centroid in ϕ mitigates this degradation ($\Delta r = +0.21$, $p < 10^{-6}$) without reducing RMS performance.

1. INTRODUCTION

Virtual orchestral instruments and professional sample libraries enable composers and producers to create realistic-sounding orchestral renderings in a digital environment. Achieving musical expressivity in such renderings critically depends on careful shaping of continuous MIDI controller signals. In particular, MIDI CC1 plays an important role: in the BBC Symphony Orchestra Core (BBC SO) library [1], CC1 simultaneously governs the dynamic level and timbral aspects by crossfading between recorded dynamic layers of different spectral character [2,3]. Automatically transcribing CC1 from real audio recordings would make it possible to transfer expressive performances to virtual instruments, and provide a compact symbolic representation of expressive dynamics for generative and analytical applications.

The fundamental challenge is the absence of ground-truth CC1 annotations in real recordings. For piano, a MIDI keyboard directly records ground-truth MIDI velocity during performance. In contrast, bowed-string dynamics arise from continuous bow motion rather than discrete keystrokes, and no equivalent recording mechanism exists. Recent work [4] demonstrated expression transcription from wind and string audio; however, the rendering pipeline (FluidSynth [5] with a General MIDI SoundFont) maps the expression controller solely to loudness, with no timbral variation. By contrast, on real bowed-string instruments, dynamic changes inherently alter timbre. Transcribing CC1 therefore recovers a richer signal encoding both loudness level and timbral character, enabling more faithful reproduction of expressive intent through BBC SO than loudness-only approaches.

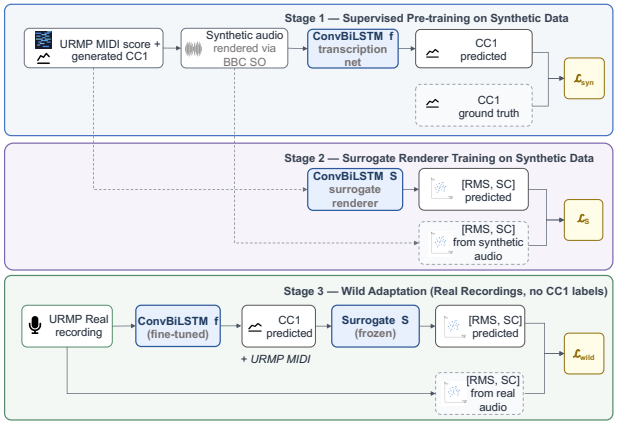


Figure 1. Overview of the proposed three-stage framework

In this work, we focus on violin as a representative bowed-string instrument and address CC1 transcription from real audio through a *differentiable analysis-by-synthesis* framework (Figure 1). Because BBC SO is proprietary and non-differentiable, we learn a lightweight *surrogate renderer* mapping an aligned MIDI score and CC1 trajectory to a compact expressive feature space $\phi = [\text{RMS}, \text{SC}]$, providing stable gradients for adaptation to real recordings. Our framework comprises three sequentially trained stages: (i) a Convolutional Bidirectional LSTM (ConvBiLSTM) transcription network pre-trained

Copyright: © 2026. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

on synthetic BBC SO renders with procedurally generated CC1 supervision; (ii) a differentiable surrogate renderer trained to predict expressive features from (score, CC1); and (iii) a wild adaptation step that fine-tunes the transcription network on each target recording by minimising the discrepancy between surrogate-predicted and observed expressive features, requiring no CC1 annotations. On 34 URMP [6] violin recordings, wild adaptation improves RMS-envelope correlation from $r = 0.675$ (synthetic-only baseline) to $r = 0.844$ on all 34 pieces ($p < 10^{-6}$, Cohen's $d = 1.90$). We further find that RMS-only adaptation can harm timbral alignment ($r_{\text{CC1-SC}} = -0.087$), while including spectral centroid in ϕ mitigates this degradation ($\Delta r = +0.21$, $p < 10^{-6}$) without reducing RMS performance.

In summary, the contributions are as follows:

- The first system, to our knowledge, for transcribing continuous expressive dynamics (CC1) from real bowed-string recordings, recovering a control signal that jointly modulates loudness level and timbral character.
- A differentiable surrogate renderer enabling analysis-by-synthesis adaptation to real recordings without CC1 annotations.
- Consistent improvement across all 34 URMP violin pieces under analysis-by-synthesis evaluation with BBC SO.

The remainder of this paper is organised as follows. Section 2 surveys related work. Section 3 formalises the task. Section 4 describes the data. Section 5 presents the proposed framework. Section 6 reports experimental results, and Section 7 concludes.

2. RELATED WORK

2.1 Expressive performance transcription

Transcribing expressive dynamics parameters from audio has been most extensively studied for piano [7–10], where MIDI recordings provide ground-truth note velocities [11, 12]. For bowed-string and wind instruments, no equivalent annotation scheme exists. The closest precursor to the present work [4] addresses expression transcription for wind and string audio via Sim2Real transfer, but is limited to loudness-based expression proxies due to its FluidSynth [5] rendering pipeline. In contrast, the present work targets BBC SO's CC1 controller, which jointly governs loudness level and timbral aspects.

2.2 Synthetic data and domain adaptation

Using rendered audio as a source of labelled training data is an established strategy [4, 13, 14] for overcoming annotation scarcity in music information retrieval. Pre-training on synthetically rendered audio has been shown to support generalisation to real recordings in expression transcription [4]. Our work extends this strategy by combining synthetic pre-training with differentiable analysis-by-synthesis adaptation to unlabelled real recordings, without requiring any CC1 annotations.

2.3 Differentiable synthesis and analysis-by-synthesis

Differentiable digital signal processing (DDSP) [15] demonstrated that embedding differentiable synthesisers into neural networks enables gradient-based estimation of acoustic parameters such as pitch and timbre without direct supervision, establishing a class of analysis-by-synthesis methods in which a differentiable forward model is optimised to match real audio observations. Such approaches typically assume a differentiable renderer is available or operate in the full waveform domain. Our work extends this paradigm to a non-differentiable, proprietary sample library by learning a surrogate renderer in a compact expressive feature space, enabling gradient-based CC1 adaptation without access to renderer internals.

3. TASK DEFINITION

We consider the task of transcribing continuous expressive dynamics directly from audio. Given an audio recording x, our goal is to estimate a continuous CC1 control trajectory via a sequence-to-sequence operator f:

$$\hat{c} = f(x), \tag{1}$$

where $\hat{c}$ denotes the predicted time-varying CC1 control trajectory. This is an *instance-specific* task: the goal is to recover the CC1 trajectory for a given, known recording; applying the system to a new recording requires only re-running Stage 3 adaptation with that recording as the target. By design, adaptation on the target recording is part of inference, not model selection. Evaluation therefore measures per-instance recovery accuracy rather than cross-piece generalisation, which is consistent with this task formulation.

4. DATA

4.1 Source Dataset: URMP

We use the University of Rochester Multi-modal Music Performance (URMP) dataset [6, 16] as the sole data source. URMP provides isolated recordings of individual instruments in ensemble performances, together with aligned MIDI scores derived from the sheet music. We focus exclusively on violin tracks, yielding 34 recordings across a variety of repertoire. Crucially, URMP contains no expressive control annotations (e.g. CC1): it provides pitch and timing information only. We exploit the dataset in two complementary ways, described below.

4.2 Synthetic Data Generation

While MID-FiLD [17] includes fine-grained CC1 dynamics data, these annotations are confounded by concurrent changes in velocity and CC11 (Expression, a secondary gain controller that typically operates independently of timbral character). Such interdependencies prevent clear disentanglement of CC1's specific impact on the audio, rendering these mixed annotations unsuitable for training a targeted, single-controller transcription model. We therefore construct a supervised training set from URMP scores

and procedurally generated CC1 curves, rendered through BBC SO with CC1 as the *sole* expressive variable.

4.2.0.1 CC1 trajectory synthesis

For each of the 34 violin pieces, we generate $K = 40$ independent CC1 trajectories using a constrained random walk: controller values evolve via smooth incremental updates, with step variance reduced near the MIDI boundaries $[0, 127]$ to avoid saturation and occasional larger jumps for diversity. CC1 is set to zero during rests. This yields $34 \times 40 = 1{,}360$ candidate trajectories.

4.2.0.2 Audio rendering

Each (score, CC1) pair is rendered via BBC SO, producing synthetic training triplets (x_s, M_s, c_s) (audio signal, MIDI score, and CC1 trajectory), all precisely aligned by construction.

4.3 Real Recordings

The original 34 URMP violin recordings provide pairs (x_w, M_w): real audio with aligned scores but *no* CC1 annotations. These are used for two purposes: (i) wild adaptation of the transcription model (Stage 3), and (ii) analysis-by-synthesis evaluation (Section 6.4). These recordings were captured in controlled acoustic conditions and represent genuine human performances, providing a realistic domain gap with respect to the synthetic training data.

5. METHOD

Our framework comprises three sequentially trained stages (Figure 1). Stage 1 pre-trains a ConvBiLSTM transcription network on synthetic BBC SO renders with full CC1 supervision. Stage 2 trains a lightweight surrogate renderer to approximate the CC1-to-expressive-feature mapping of BBC SO in a compact feature space $\boldsymbol{\phi} = [\phi_{\mathrm{RMS}}, \phi_{\mathrm{SC}}]$. Stage 3 fine-tunes the transcription network on unlabelled real recordings by minimising the discrepancy between surrogate-predicted and observed expressive features, requiring no CC1 annotations.

5.1 Stage 1: CC1 Transcription Model

We model CC1 transcription as a regression problem from audio to a continuous expressive control trajectory (Section 3), with $\hat{c}$ representing the predicted CC1 trajectory.

5.1.0.1 Audio representation

The input audio x is converted into a logmel spectrogram, then concatenated with a z-normalised frame-level RMS feature. The resulting feature representation is denoted by $\mathbf{X}$.

5.1.0.2 Model architecture

The transcription function $f(\cdot)$ is implemented as a ConvBiLSTM network. Two Conv1d layers (128 channels, kernel size 5, ReLU activations) extract local patterns from the input tensor $\mathbf{X}$, followed by a 2-layer bidirectional LSTM (256 hidden units per direction) that captures long-range temporal dependencies. A final linear projection maps the BiLSTM output to a scalar at each frame, producing the predicted CC1 trajectory $\hat{c}$.

5.1.0.3 Training objective

The model is trained with a mean-squared error loss augmented by a first-difference penalty encouraging smooth trajectories:

$$\mathcal{L}_{\text{syn}} = \mathrm{MSE}(\hat{c}_s, c_s) + \lambda_{\text{slope}} \mathrm{MSE}(\Delta\hat{c}_s, \Delta c_s)\,, \qquad (2)$$

where Δ denotes the first-difference operator applied frame-wise and λ_{slope} is a scalar weight.

5.2 Stage 2: Surrogate Renderer

Since BBC SO is non-differentiable, we introduce a differentiable surrogate renderer. The surrogate learns the mapping from an aligned MIDI score M and a CC1 trajectory c to a low-dimensional feature sequence, providing stable gradients for analysis-by-synthesis adaptation.

Given an aligned MIDI score M and a CC1 control trajectory c, the surrogate renderer predicts a sequence of expressive audio features

$$\hat{\phi} = S(M, c), \qquad (3)$$

where $\hat{\phi}$ denotes the predicted frame-level expressive features.

5.2.0.1 Expressive feature representation

We define $\boldsymbol{\phi}$ as a two-dimensional frame-level feature vector:

$$\boldsymbol{\phi} = \big[\phi_{\mathrm{RMS}},\ \phi_{\mathrm{SC}}\big], \qquad (4)$$

where ϕ_{RMS} is the frame-level RMS amplitude, capturing the dynamic envelope most directly controlled by CC1, and ϕ_{SC} is the spectral centroid (SC), capturing the timbral brightness that CC1 simultaneously modulates by crossfading between recorded dynamic layers of different spectral character. ϕ_{RMS} is the primary target, while ϕ_{SC} acts as a timbral correlate whose tracking we treat as a secondary quality criterion (Section 6.4.0.4). While numerous timbral descriptors exist, we use the spectral centroid as it has been established as a robust acoustical correlate of expression in sustained-pitch instruments [18, 19]. Both dimensions are z-normalised independently for numerical stability.

5.2.0.2 Surrogate architecture

The surrogate renderer S is implemented as a ConvBiLSTM network (same structure as f) with a 3-dimensional input (CC1 z-score, normalised pitch, note-active flag) and a 2-dimensional output $[\mathrm{RMS},\ \mathrm{SC}]$. The note-active flag is a binary indicator (1 during pitched notes, 0 during rests) that informs the model when to apply expressive modulation and when to expect silence.

5.2.0.3 *Surrogate model training*

The surrogate renderer $S(\cdot)$ is trained using synthetic data consisting of triplets (x_s, M_s, c_s) with temporally aligned content. Let $\boldsymbol{\phi}(x_s)$ denote the two-dimensional expressive feature sequence extracted from the synthetic audio x_s. The surrogate renderer is trained to minimise the regression loss

$$\mathcal{L}_{\text{S}} = \left\| S(M_s, c_s) - \boldsymbol{\phi}(x_s) \right\|_1 . \tag{5}$$

After training, the surrogate renderer provides a differentiable approximation of how CC1 control trajectories affect expressive audio characteristics under the synthetic rendering condition.

5.3 Stage 3: Wild Adaptation via Analysis-by-Synthesis

In the real-recording setting, no ground-truth CC1 annotations are available. We therefore adopt a differentiable analysis-by-synthesis framework operating in the expressive feature domain. Given a real audio recording x_w and its aligned MIDI score M_w, the transcription model first predicts a CC1 trajectory

$$\hat{c}_w = f(x_w), \tag{6}$$

where $\hat{c}_w$ denotes the predicted CC1 trajectory. The predicted control trajectory is then passed, together with the MIDI score, to the surrogate renderer to obtain predicted expressive features

$$\hat{\phi}_w = S(M_w, \hat{c}_w). \tag{7}$$

In parallel, expressive features $\phi(x_w)$ are extracted directly from the real audio recording. We enforce expressive consistency by minimising the discrepancy between the predicted and observed expressive features

$$\mathcal{L}_{\text{wild}} = \left\| \hat{\phi}_w - \phi(x_w) \right\|_1 , \tag{8}$$

where $\phi(x_w)$ denotes the expressive feature sequence extracted from x_w.

By optimising this objective, the transcription model is encouraged to produce CC1 trajectories whose induced expressive characteristics are consistent with those observed in real performances, despite the absence of explicit CC1 annotations.

5.4 Training Procedure

The three stages are trained sequentially. Stage 1 trains f on synthetic data using $\mathcal{L}_{\text{syn}}$. Stage 2 trains S independently on the same synthetic set using $\mathcal{L}_{\text{S}}$. Stage 3 holds S fixed and fine-tunes f using a combined objective

$$\mathcal{L} = \mathcal{L}_{\text{syn}} + \alpha \, \mathcal{L}_{\text{wild}}, \tag{9}$$

where α controls the relative weight of the real-recording adaptation objective. The inclusion of $\mathcal{L}_{\text{syn}}$ in Stage 3 is deliberate: optimising $\mathcal{L}_{\text{wild}}$ alone on real data can cause the model to lose its previously learned CC1 semantics (catastrophic forgetting). By simultaneously drawing one synthetic mini-batch (contributing $\mathcal{L}_{\text{syn}}$, with labelled CC1) and one real mini-batch (contributing $\mathcal{L}_{\text{wild}}$, unlabelled) at each training step, the model retains its synthetic CC1 knowledge while gradually adapting its predictions to the real recording domain.

6. EXPERIMENTS

6.1 Experimental Setup

6.1.0.1 *Datasets*

The synthetic training set consists of 1,224 triplets (x_s, M, c_s) (90% of the 1,360 generated; 10% held out for testing): audio rendered with the BBC SO library, the MIDI score M, and a procedurally generated CC1 trajectory c_s. For wild adaptation and evaluation we use the original 34 URMP violin recordings x_w with aligned scores M_w but no CC1 annotations. Consistent with the instance-specific nature of the task (Section 3), the same recordings are used for both adaptation and evaluation.

6.1.0.2 *Features*

Audio is represented as a 128-band log-mel spectrogram ($f_s = 22{,}050$ Hz, $N_{\text{FFT}} = 2{,}048$, hop $= 512$), concatenated with a z-normalised RMS feature. CC1 values are z-normalised across the training set for learning; CC1 predictions are recovered to the standard 0-127 MIDI range by the inverse z-score transform at inference time.

6.1.0.3 *Models*

The transcription model f is a ConvBiLSTM: two Conv1d layers (128 channels, kernel size 5) followed by a 2-layer BiLSTM (256 hidden units per direction) and a linear output head. The surrogate renderer S uses the same ConvBiLSTM architecture as f, but with a 3-dimensional input (CC1 z-score, normalised pitch, note-active flag) and a 2-dimensional output [RMS, SC].

6.1.0.4 *Training details*

Stage 1 is trained for 50 epochs on the synthetic set with the AdamW optimiser (learning rate lr $= 10^{-3}$, 15% validation split), using the objective in Eq. (2) with $\lambda_{\text{slope}} = 0.1$. Stage 2 (surrogate renderer) is trained for 100 epochs on the same set (lr $= 10^{-3}$, 20% validation split). Stage 3 (wild adaptation) fine-tunes f for 20 epochs using $\alpha = 0.1$ (lr $= 10^{-4}$, Adam), with S held fixed.

6.2 In-Domain Validation (Stage 1)

On the held-out synthetic validation split, the Stage 1 model achieves a frame-level MAE of 0.095 and MSE of 0.018 (z-score space), confirming that it can reliably recover expressive CC1 trajectories from rendered audio.

However, the domain gap between BBC SO-rendered audio and real performances may limit generalisation; Stage 3 wild adaptation is designed to bridge this gap, as evaluated next.

6.3 Surrogate Renderer Validation (Stage 2)

The surrogate renderer S is trained to predict frame-level [RMS, SC] from (CC1 z-score, normalised pitch, note-active flag) using the synthetic training set. On the held-out validation split, the RMS head achieves a MAE of 0.089 in z-score space, indicating that S adequately captures the CC1-driven dynamic mapping needed for adaptation; SC behaviour is further analysed in Section 6.4.0.4.

6.4 Analysis-by-Synthesis Evaluation on Real Recordings (Stage 3)

Since real recordings carry no ground-truth CC1 labels, direct evaluation of transcription accuracy is not possible. Instead, we adopt an analysis-by-synthesis protocol: predicted CC1 trajectories are rendered into audio via BBC SO and compared against the original URMP recordings. This provides an external assessment of expressive fidelity using the actual target synthesiser rather than the learned surrogate renderer.

6.4.0.1 Protocol

For each of the 34 URMP violin pieces and each experimental condition, we generate a MIDI file in which the note sequence is derived from the aligned score and the CC1 automation is set according to that condition. Each MIDI file is then rendered to audio using BBC SO, producing a synthetic performance $\hat{x}$. We compare the RMS envelope of $\hat{x}$ against that of the original URMP recording x_w using Pearson correlation r as the primary measure of dynamic fidelity. Critically, this evaluation uses the actual BBC SO synthesiser, a non-differentiable, proprietary renderer entirely separate from the learned surrogate S, providing an independent external assessment. Although Stage 3 adaptation minimises RMS discrepancy via the surrogate S, S is an imperfect approximation of BBC SO trained on a finite synthetic corpus; the final RMS correlation is measured against the true synthesiser, not S itself. Improvements in BBC SO RMS correlation therefore confirm that adaptation benefits transfer beyond surrogate fitting, yielding CC1 trajectories that are more consistent with the dynamic behaviour of the actual renderer.

6.4.0.2 Conditions

We evaluate four conditions:

- **Constant CC1**: CC1 fixed at the mid-range value (MIDI 64), establishing a note-structure-only baseline.
- **Stage 1**: CC1 predicted by the synthetically trained model (no wild adaptation); this condition corresponds to the approach of [4].
- **Stage 3 (RMS)**: wild adaptation with $\phi = [\text{RMS}]$ only (ablation, single-objective).
- **Stage 3 (RMS+SC)**: wild adaptation with $\phi = [\text{RMS}, \text{SC}]$ (proposed, two-objective).

6.4.0.3 RMS results

Table 1 summarises the analysis-by-synthesis evaluation. Both Stage 3 variants (RMS and RMS+SC) achieve near-identical RMS correlation ($r = 0.845$ and $r = 0.844$ respectively), substantially outperforming Stage 1 ($r = 0.675$) and the constant baseline ($r = 0.403$). Wild adaptation also reduces the standard deviation of r from 0.121 (Stage 1) to 0.037–0.045, indicating more consistent performance across pieces.

Statistical significance was assessed using the Wilcoxon signed-rank test. Stage 3 (RMS+SC) vs. Stage 1: $p = 1.82 \times 10^{-7}$, $d = 1.90$, all 34 pieces improved. Stage 3 (RMS+SC) vs. Stage 3 (RMS): $p = 0.585$ (not significant), confirming that adding SC to the adaptation objective does not degrade RMS tracking. The improvement from Stage 1 to Stage 3 (RMS+SC) ($\Delta r = +0.169$) represents genuine dynamic information recovered through wild adaptation.

Table 1. Analysis-by-synthesis RMS evaluation on 34 URMP violin recordings. r: RMS envelope Pearson correlation between BBC SO rendered audio and real recording (higher is better).

Condition	mean r	std r
Constant CC1	0.403	0.117
Stage 1	0.675	0.121
Stage 3 (RMS)	0.845	0.037
Stage 3 (RMS+SC)	0.844	0.045

6.4.0.4 Spectral centroid results

To probe the effect of CC1 on timbral quality, we compute the Pearson correlation between the predicted CC1 trajectory and the real recording's spectral centroid ($r_{\text{CC1-SC}}$). Table 2 reports these values. RMS-only adaptation (Stage 3/RMS) causes the CC1–SC correlation to collapse to $r = -0.087$, indicating that the adapted model's dynamics are anti-correlated with real timbral brightness. The two-objective adaptation (Stage 3/RMS+SC) mitigates this degradation, recovering $r = +0.125$ — a highly significant improvement over Stage 3 (RMS) ($\Delta r = +0.21$, Wilcoxon $p = 1.82 \times 10^{-7}$, $d = 1.58$, all 34 pieces improved).

Table 2. Spectral centroid (SC) tracking: Pearson correlation between predicted CC1 trajectory and real recording's SC ($r_{\text{CC1-SC}}$), probing the timbral alignment of the adapted CC1 trajectory.

Condition	mean $r_{\text{CC1-SC}}$	std
Constant CC1	n/a†	—
Stage 1	$+0.262$	0.157
Stage 3 (RMS)	-0.087	0.236
Stage 3 (RMS+SC)	+0.125	0.235

†Constant CC1 has zero variance; Pearson r is undefined.

6.5 Qualitative Analysis

Figure 2 shows mean-centred RMS envelopes for a representative 5-second excerpt of *Jupiter* (a URMP violin piece, 44–49 s). Stage 3 (red) tracks the dynamic contour of the original recording (grey) closely, with peaks

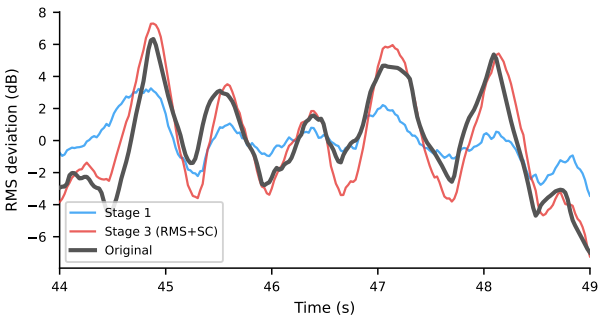


Figure 2. Mean-centred RMS envelopes for a 5-second excerpt of *Jupiter* (a URMP violin piece, AuSep_1_vn_01), 44–49 s. Stage 3 (red) closely follows the original's dynamic contour (grey); Stage 1 (blue) is markedly flatter.

and valleys well aligned throughout the excerpt. Stage 1 (blue) produces a substantially flatter envelope, failing to reproduce the pronounced dynamic swings present in the real performance. This qualitative observation is consistent with the quantitative results in Table 1: wild adaptation enables the model to recover expressive dynamics that are invisible from synthetic pre-training alone.

7. CONCLUSION

We presented a differentiable analysis-by-synthesis framework for transcribing continuous expressive dynamics from real bowed-string audio, using violin as a case study. The key enabling component is a learned surrogate renderer that provides a differentiable path in a compact expressive feature space, allowing adaptation to unlabelled real recordings without any CC1 annotations. Evaluation via analysis-by-synthesis reveals that wild adaptation improves performance. This improvement over the synthetic-only baseline validates the effectiveness of incorporating real-world recordings into the training pipeline. We further showed that RMS-only adaptation can harm timbral alignment, and that including spectral centroid in the adaptation objective mitigates this degradation without sacrificing RMS performance.

Future work could explore replacing the aligned MIDI score with automatic pitch tracking to enable score-free adaptation, and testing surrogate generalisation to held-out pieces or alternative virtual instrument libraries.

Acknowledgments

This work is supported by the EPSRC UKRI Centre for Doctoral Training in Artificial Intelligence and Music (Grant no. EP/S022694/1). We would also like to express our sincere thanks to our industry partner, Spitfire Audio, for their continued support and for providing the BBC Symphony Orchestra Core plugin.

8. REFERENCES

[1] Spitfire Audio, "BBC Symphony Orchestra Core," 2025, virtual Instrument Library. [Online]. Available: https://www.spitfireaudio.com/bbc-symphony-orchestra-core

[2] Universal Audio Help Center, "Spitfire Audio—Controllers," 2022. [Online]. Available: https://help.uaudio.com/hc/en-us/articles/360041911651-Spitfire-Audio

[3] Spitfire Audio, "How to create orchestra music with virtual instruments: Tip 6 Use controllers," 2025. [Online]. Available: https://www.spitfireaudio.com/blogs/blog/how-to-create-orchestra-music-with-virtual-instruments-a-beginners-guide

[4] Y. Xie, Z. Guo, and M. Barthet, "Fine-Grained MIDI Expression Transcription from Wind and String Instrument Audio via Sim2Real Transfer Learning ," in *Proceedings of the 17th International Symposium on Computer Music Multidisciplinary Research (CMMR)* , 2025, pp. 570–581.

[5] "FluidSynth," [Online]. Available: https://www.fluidsynth.org/ (accessed July 14, 2026).

[6] "University of Rochester Multi-Modal Music Performance (URMP) Dataset," [Online]. Available: https://labsites.rochester.edu/air/projects/URMP.html, 2018.

[7] Z. He, D. Cooper, D. Huang, and R. Togneri, "Filling MIDI Velocity using U-Net Image Colorizer," in *Proceedings of the 17th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2025.

[8] H. Kim and X. Serra, "A Method for MIDI Velocity Estimation for Piano Performance by a U-Net with Attention and FiLM," in *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[9] H. Kim, M. Miron, and X. Serra, "Score-Informed MIDI Velocity Estimation for Piano Performance by FiLM Conditioning," in *Proceedings of the 20th Sound and Music Computing Conference (SMC)*, Stockholm, Sweden, 2023.

[10] D. Jeong, T. Kwon, Y. Kim, and J. Nam, "VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance," in *The 20th International Society for Music Information Retrieval Conference (ISMIR)*. International Society for Music Information Retrieval Conference (ISMIR), 2019, pp. 908–915.

[11] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in *International Conference on Learning Representations (ICLR)*, 2019.

[12] M. Müller, V. Konz, W. Bogler, and V. Arifi-Müller, "Saarland Music Data (SMD)," in *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011.

[13] S. Sarkar, E. Benetos, and M. Sandler, "EnsembleSet: A New High-Quality Synthesised Dataset for Chamber Ensemble Separation," in *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[14] F. Ostermann, I. Vatolkin, and M. Ebeling, "AAM: A Dataset of Artificial Audio Multitracks for Diverse Music Information Retrieval Tasks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 13, 2023.

[15] J. Engel, L. Hantrakul, C. Gu, and A. Roberts, "DDSP: Differentiable Digital Signal Processing," in *International Conference on Learning Representations (ICLR)*, 2020.

[16] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2018.

[17] J. Ryu, S. Rhyu, H.-G. Yoon, E. Kim, J. Y. Yang, and T. Kim, "MID-FiLD: MIDI dataset for fine-level dynamics," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 1, 2024, pp. 222–230.

[18] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad, "Acoustical correlates of timbre and expressiveness in clarinet performance," *Music Perception*, vol. 28, no. 2, pp. 135–154, 2010.

[19] ——, "Analysis-by-synthesis of timbre, timing, and dynamics in expressive clarinet performance," *Music Perception*, vol. 28, no. 3, pp. 265–278, 2011.