View Reviews

Paper ID
3105

Paper Title
ViolinDiff: Enhancing Expressive Violin Synthesis with Pitch Bend Conditioning

Track Name
ICASSP 2025 Main Tracks

Reviewer #1

Questions

2. Importance/Relevance
3. Of sufficient interest

3. Justification of Importance/Relevance Score (required if score is 1 or 2).
To my knowledge, it is the first explicit investigation into violin synthesis with diffusion models. See "Additional comments" (20.) for more details.

5. Originality/Novelty
3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)
See "Additional comments" (20.) for details.

7. Theoretical Development
3. Probably correct; provides limited new insights or understanding

9. Experimental Validation
2. Lacking in some respect

10. Justification of Experimental Validation Score (required if score is 1 or 2).
See "Additional comments" (20.) for details.

11. Clarity of Presentation
3. Clear enough

13. Reference to Prior Work
3. References adequate

15. Overall evaluation of this paper
4. Definite accept

16. Justification of Overall evaluation of this paper (required)
See "Additional comments" (20.) for details.

20. Additional comments to author(s): (Required if no other justification comments have been provided above.)
The paper introduces a diffusion-based approach for synthesizing expressive Violin performance signals from MIDI. The proposed model contains a two-step approach of first generating an expressive F0 trajectory representation (similar to state-of-the-art approaches in singing voice synthesis, but with the ability to represent polyphony), and then synthesizing a mel spectrogram using the F0 trajectory as input. The approach is evaluated with some FAD calculations, a note-wise comparison of vibrato presence in real and synthesized audio, and a listening test.

From my point of view, this is a relevant and sufficiently novel contribution to be published at ICASSP. To my knowledge, it is the first explicit investigation into violin synthesis with diffusion models. However, particularly the evaluation of the approach has some weaknesses and shortcomings (see below), which could partly be addressed in a minor revision. With further possibilities for improving the method and evaluation metrics in the future, the paper has the potential for triggering a productive discussion at the conference and beyond.

Some notes and questions for a potential revision:

(1) Additional baselines would be very useful to understand the relative scale of the objective metrics in Table 1. For example, one could include a commercial MIDI solo violin synthesis (e.g., Vienna String Library or Samplemodeling) to show how big the FAD and Vibrato F1/MAE differences are to such renderings (either from plain MIDI or with manual programming for a high-quality baseline). Without such a reference, the differences between the proposed approach and the "NoBend" condition look quite small across all FAD categories. An F1 score for vibrato presence below 0.6 also does not seem very good in isolation, since - if I understand correctly - it would imply that only half the notes have an appropriate amount of vibrato compared to real performances?

(2) Similarly, why are the models by Hawthorne and Maman used in the listening test but not included in the objective comparisons in Table 1? Another possible baseline could have been the mentioned MIDI-DDSP.

(3) It would be interesting to test how well the synthesis model adheres to the R_bend input, e.g., by providing synthetic inputs with controlled bends and vibratos.

(4) The explicit focus on vibrato and the binary decision of its presence (in both F1 score and Perf-MAE, if I understand correctly) seems to be a very indirect way of measuring the

model performance. I understand that it is not possible to compare F0 trajectories directly due to the generative nature of the "Bend Estimation Module", but maybe it would be insightful to characterize the F0 trajectories of real and generated examples in terms of their note-wise statistics (e.g., mean and standard deviation from the 12-TET pitch) ro reveal if the proposed system behaves comparably to real performers. (In contrast to a statement in the paper, I would not expect that F0 estimation error plays a significant role in such an analysis, particularly when considering monophonic passages.)

(5) I would expect the amount of vibrato and other pitch deviations to depend strongly on the performer. An analysis of model output differences depending on e_p would be interesting.

(6) Up to some ambiguity due to the range choice of -1 to 1 instead of -0.5 to 0.5, R_bend contains redundant information compared to R_frame. I wonder if this could be used to simplify the input representations somehow. On the other hand, redundancy exists between all elements in $\mathcal{R}$, and the redundancy of frame and onset representation has been proven useful in previous work...

(7) Minor issues:
* Fig. 2: The differences are very hard to see. There is much more space available, so it would be good to use it and further zoom in on the region of interest.
* In the introduction, typical MOS for diffusion-based singing and instrument synthesis results is given using the 5 point scale recommended by ITU-T P.800. The listening test in this paper uses a 100-point scale instead, so that it is even more difficult to put the results in relation to previous approaches.
* Without familarity with the cited references, section IV.A is quite hard to understand.
* Page 1: Extra space in second to last line of the first column

Reviewer #2

Questions

2. Importance/Relevance

4. Of broad interest

5. Originality/Novelty

3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)

The paper introduces ViolinDiff, a two-stage diffusion-based model designed to estimate the fine-granular F0 contour in the initial stage, followed by mel-spectrogram generation in the second stage for synthesizing violin music transcriptions. This structured approach is

both novel and well-justified, effectively capturing intricate pitch variations essential for violin timbre. The resulting synthesized sound quality is notably impressive, indicating a high potential for advancements in realistic violin synthesis.

7. Theoretical Development

3. Probably correct; provides limited new insights or understanding

8. Justification of Theoretical Development Score (required if score is 1 or 2).

While I am not a specialist in diffusion-based models, I found that the paper's use of foundational model components, clear conceptual presentation, and well-structured mathematical formulations make the theoretical development accessible and comprehensible. The authors have successfully organized complex concepts in a way that facilitates understanding, even for readers less familiar with this modeling approach.

9. Experimental Validation

3. Limited but convincing

10. Justification of Experimental Validation Score (required if score is 1 or 2).

The paper is mostly based on empirical results, rather than being a theoretical paper.

11. Clarity of Presentation

4. Very clear

12. Justification of Clarity of Presentation Score (required if score is 1 or 2).

The paper presents its ideas, contributions, and limitations with a very high clarity. Additionally, the fair comparison with other baseline models, particularly in the Results section, further enhances its transparency. This thorough and well-structured presentation strengthens the overall quality of the paper.

13. Reference to Prior Work

4. Excellent references

15. Overall evaluation of this paper

4. Definite accept

16. Justification of Overall evaluation of this paper (required)

This paper presents ViolinDiff, a novel two-stage diffusion-based model for synthesizing violin music with fine-granular F0 contour estimation followed by mel-spectrogram generation. The approach is innovative and effective, with impressive sound quality in the synthesized recordings, demonstrating a high potential for advancing realistic violin synthesis.

The theoretical development is clearly articulated, with well-defined components, conceptual clarity, and accessible mathematical formulation. This makes the paper understandable, even to those less familiar with diffusion-based models, showcasing the authors' ability to present complex concepts effectively.

As a final suggestion, I would recommend placing less emphasis on the contribution to

polyphonic music, as violin music is not primarily known for polyphony. Instead, the paper's strength lies in its synthesis based on fine-granular pitch bends. Furthermore, it could be interesting to explore an alternative approach in Section V.D.2, aligning audio recordings of different performances and using anchor points to interpolate their transcriptions.

Reviewer #3

Questions

2. Importance/Relevance

4. Of broad interest

5. Originality/Novelty

3. Moderately original; provides limited new insights or understanding

6. Justification of Originality/Novelty Score (required)

This paper proposed ViolinDiff, a diffusion-based MIDI-to-audio synthesis model that predicts polyphonic pitch bend and leverages it to synthesize expressive violin performances.

7. Theoretical Development

3. Probably correct; provides limited new insights or understanding

9. Experimental Validation

3. Limited but convincing

11. Clarity of Presentation

4. Very clear

13. Reference to Prior Work

4. Excellent references

15. Overall evaluation of this paper

3. Marginal accept

16. Justification of Overall evaluation of this paper (required)

This paper is well written.
It proposed ViolinDiff, a diffusion-based MIDI-to-audio synthesis model that predicts polyphonic pitch bend and leverages it to synthesize expressive violin performances.