ICASSP 2022

2022 IEEE International Conference on Acoustics, Speech and Signal Processing, 22-27 May 2022, Singapore

Author Rebuttal to Paper Reviewers

Upload your reviewer rebuttal by dragging the **ONE-PAGE PDF** file onto the **Reviewer Rebuttal (PDF)** boxes below.

Only one file to each reviewer may be uploaded. If you upload another rebuttal to a reviewer, the newer upload will override the older upload.

Paper Information

Paper Number

2805

Paper Title

DEEP PERFORMER: SCORE-TO-AUDIO MUSIC PERFORMANCE SYNTHESIS

Authors

Hao-Wen Dong, University of California San Diego
Cong Zhou, Dolby Laboratories
Taylor Berg-Kirkpatrick, University of California San Diego
Julian McAuley, University of California San Diego

Reviewers

Reviewer 2968

Importance/Relevance to ICASSP 2022 →	Of sufficient interest
Justification of Importance/Relevance Score	
[None Provided by Reviewer]	
Novelty/Originality →	Moderately original
Justification of Novelty/Originality Score	

Upload Rebuttal to Reviewer 2968

To upload, drag the one-page PDF file of your rebuttal to review 2968 . If you cannot

[None Provided by Reviewer]

Technical Correctness → Probably correct

Justification of Technical Correctness Score

[None Provided by Reviewer]

Experimental Validation → Sufficient validation/theoretical paper

Justification of Experimental Validation Score

[None Provided by Reviewer]

Clarity of Presentation → Clear enough

Justification of Clarity of Presentation Score

[None Provided by Reviewer]

Reference to Prior Work → References adequate

Justification of Reference to Prior Work Score

[None Provided by Reviewer]

Additional comments to author(s)

This paper presents a novel system for score-to-audio music performance synthesis. It consists of three stages: (1) an alignment model, (2) a synthesis model, and (3) an inversion model.

The manuscript is well written and organized. The authors wittily describe the advantages and weaknesses of their proposal. The experimentation is quite illustrative and reveals promising results.

For further experimentation, I recommend using the URMP dataset:

Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2018). Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Transactions on Multimedia, 21(2), 522-535.

Moreover, the authors release the dataset and the source code for the alignment process to the public, which is very valuable to facilitate future research on score-to-audio music synthesis.

I recommend accepting the paper for publication.

drag a file on this box, you can also click anywhere in this box to open a file chooser.

IMPORTANT: The rebuttal must **ONLY** respond to reviewer 2968

Uploading a rebuttal is optional to each reviewer.

Reviewer 0327

Importance/Relevance to ICASSP 2022 → Of sufficient interest

Justification of Importance/Relevance Score

[None Provided by Reviewer]

Novelty/Originality → Moderately original

Justification of Novelty/Originality Score

[None Provided by Reviewer]

Upload Rebuttal to Reviewer 0327

To upload, drag the one-page PDF file of your rebuttal to review 0327 . If you cannot

Technical Correctness → Probably correct

Justification of Technical Correctness Score

[None Provided by Reviewer]

Experimental Validation → Limited but convincing

Justification of Experimental Validation Score

[None Provided by Reviewer]

Clarity of Presentation → Clear enough

Justification of Clarity of Presentation Score

[None Provided by Reviewer]

Reference to Prior Work → References adequate

Justification of Reference to Prior Work Score

[None Provided by Reviewer]

Additional comments to author(s)

The authors present a deep learning based score-to-audio synthesizer and evaluate it with violin and piano data. The paper is generally well written and easy to understand, albeit values of key hyper parameters, such as the dimensions of the employed embeddings for example, are missing and should be provided. The proposed model is interesting in that it proposes a polyphonic synthesis model. The experimental results, however, in particular the MOS results do not indicate a sound superiority over the baseline method, except for the piano case. Disconcerting in the piano case, in turn, is that the ablation study suggests that the performance of the proposed system improves when a key module of the system (note-wise positional encoding) is removed. It seems as if quite a bit further research is still necessary to imbue the system with a broader/convincing performance edge. Nevertheless, the paper is in its current form still of interest to the community and should thus be perceived.

drag a file on this box, you can also click anywhere in this box to open a file chooser.

IMPORTANT: The rebuttal must **ONLY** respond to reviewer 0327

Uploading a rebuttal is optional to each reviewer.

Reviewer 282E

Importance/Relevance to ICASSP 2022 → Of sufficient interest

Justification of Importance/Relevance Score

[None Provided by Reviewer]

Novelty/Originality → Moderately original

Justification of Novelty/Originality Score

[None Provided by Reviewer]

Technical Correctness → Probably correct

Justification of Technical Correctness Score

[None Provided by Reviewer]

Experimental Validation → Sufficient validation/theoretical paper

Justification of Experimental Validation Score

Upload Rebuttal to Reviewer 282E

To upload, drag the one-page PDF file of your rebuttal to review 282E . If you cannot drag a file on this box, you can also click anywhere in this box to open a file chooser.

IMPORTANT: The rebuttal must **ONLY** respond to reviewer 282E

[None Provided by Reviewer]

Clarity of Presentation → Very clear

Justification of Clarity of Presentation Score

[None Provided by Reviewer]

Reference to Prior Work → Excellent references

Justification of Reference to Prior Work Score

[None Provided by Reviewer]

Additional comments to author(s)

The authors describe a system for synthesizing a musical performance from a score. It utilizes Transformers and recent work in text to speech synthesis for this purpose. The paper is well written and the system is clearly described. The polyphonic mixer and note-wise positional encoding are nice ideas for handling the challenges of polyphony and note dynamics respectively. A new dataset released by the authors as well as existing datasets are used to evaluate the system. The results are evaluated and compared to a baseline system using both quantitative (MSE) and qualitative (MOS from user study) as well as some ablation experiments.

I think the paper will make a nice contribution to the increasing literature on performance rendering.

Uploading a rebuttal is optional to each reviewer.

Finish and Logout