View Reviews

Paper ID
293

Paper Title
Towards Automatic Instrumentation by Learning to Separate Parts in Multitrack Music

Track Name
Papers

Reviewer #1

Questions

2. The title and abstract reflect the content of the paper.
Strongly Agree

3. The paper discusses, cites and compares with all relevant related work.
Strongly Agree

4. The writing and language are clear and structured in a logical manner.
Strongly Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Strongly Agree

7. The content is scientifically correct.
Strongly Agree

8. The paper provides novel methods, findings or results.
Strongly Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Strongly Agree

10. The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Agree

11. Please explain your assessment of reusable insights in the paper.
The authors include well-organized and well-documented code ensuring easy reproducibility, as well as providing the reader with the convenience of building off the paper's results.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Agree

17. Overall evaluation
Strong Accept

18. Main review and comments for the authors
This paper is well-written and easy to follow. The authors convincingly evaluate the proposed models against several baselines, while also providing some insight into how the input features and data augmentation contribute to their success across several types of datasets.

If I have any criticism, it's that the paper is strongly biased toward Western music traditions. For instance, the assumption of Western pitch classes is baked into the paper. While it's clear that the proposed models are general enough to substitute a different notion of pitch classes, it would be nice for the authors to address this issue explicitly.

It would also be helpful to discuss how including gestural information, such as MIDI velocity, might change the paper's outcomes.

Some minor edits and questions follow.

Models:
165: "frequency" -> "fundamental frequency"
165: Why do the authors consider it useful to include pitch __and__ fundamental frequency as input features?
187 "that" -> "that is"
214, equation: what does a_i represent?

Bibliography
[6]: Fix capitalization of "Voice"
[20]: Capitalize "HMM"
[24]: Capitalize "RNN"
[30]: Fix capitalization for "Transformer-GANs"
[38]: Fix capitalization for "MusPy"

Reviewer #2

Questions

2. The title and abstract reflect the content of the paper.
Strongly Agree

3. The paper discusses, cites and compares with all relevant related work.
Agree

4. The writing and language are clear and structured in a logical manner.
Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by “n” pages of references, references are well formatted). If you selected “No”, please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Strongly Agree

7. The content is scientifically correct.
Agree

8. The paper provides novel methods, findings or results.
Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Agree

10. The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Agree

11. Please explain your assessment of reusable insights in the paper.
This paper's inclusion of and analysis of the effects of multiple music genres on the task of automatic instrumentation is the majority of its novelty, but the findings included (particularly related to the effects of different time encoding and different types of "hints") could bear future research and extension.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Disagree

17. Overall evaluation
Weak Accept

18. Main review and comments for the authors
#Main Points:
This paper's task is automatic instrumentation: dynamically assigning instruments to notes in solo music during performance. Issues related to this task include incorrectly assigning notes to instruments in a polyphonic context due to lack of future information (if online) and/or misinterpreting musical voicings. Applications include real-time performance of multi-instrument music using a single keyboard (replacing the fixed zone-based instrument assignment that some keyboardists will use with a dynamic system) and offline part separation from mixture in multitrack music.

#Methodology:
The authors used databases of polyphonic (but voice-separated) MIDI files in a variety of genres (Bach chorales, string quartets, video game music, and pop music). This data is then used to train a variety of ML architectures: LSMTs, bi-directional LSTMs, and Transformer variants. Features input into these model include time, pitch, duration, frequency, beat and bar position, and additionally novel features called "entry hints" (when voices enter) and "pitch hints" (average pitch of each track). These hints gave me pause upon first reading this paper as they seem to be more than just "hints", and potentially difficult to implement in a real-time context (particularly the "entry hints"?), but the authors do discuss this in Section 7.4.

#Evaluation/Results:
Using known ground truth, the author's models are quantitatively compared to a few baseline models: automatic instrumentation via tessiturae ("zones"), a simple pitch-following algorithm, and a multilayer perceptron. Online and offline models are compared, the effects of the input features (including "hints") are noted, and additionally the authors compare varying types and levels of time encoding (e.g., raw time vs. beats and bars) and transposition augmentation (e.g., no transposition vs. at all transposition levels). In short, the LSTM models always perform better than the baseline and Transformer models, different genres respond differently to both time encoding and transposition augmentation (will not go into detail here, although it is interesting).

#Grammar/Syntax/Readability:
Minor issues, clear sentences with few grammatical issues.

#Overall/Notes:
I think that the task and how the authors tackled it was clearly presented and straightforward, the inclusion of different genres and different ways of encoding features (e.g. types of time encoding) was interesting, and the results seem useful for future research. At the outset of this paper the authors emphasized the real-time focus of their automatic instrumentation task, and I would have loved to see that followed through with (having improvised works played by a solo keyboardist be evaluated on how well they can track intended instrumentation, for example), but that could be for a later paper.

Reviewer #3

Questions

2. The title and abstract reflect the content of the paper.
Agree

3. The paper discusses, cites and compares with all relevant related work.
Agree

4. The writing and language are clear and structured in a logical manner.
Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Strongly Agree

7. The content is scientifically correct.
Agree

8. The paper provides novel methods, findings or results.
Agree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Agree

10. The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

11. Please explain your assessment of reusable insights in the paper.
Authors describe some experiments for voice separation on symbolic music, not necessarily being monophonic voices. However, their results are quite hard to extend to other fields in MIR.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Disagree

17. Overall evaluation
Weak Accept

18. Main review and comments for the authors
This works on symbolic music domain but it's hard to realize that until you reach the Introduction, probably it could be specified better on the abstract as the word "Multitrack" is very common for raw audio domain.

The main case described in the paper is to allow a keyboard player have instrumentation automatically. However, the training dataset are based mostly on bigger ensembles that a single keyboard player couldn't really play. Which brings to the next point: If it is supposed to be played by a single pianist there shouldn't be cases where the same note is played by different instruments simultaneously, something that happens as it can be seen on Figure 4 and 5. It is still relevant however for automatic instrumentation cases of "downmixed" piano rolls.

As the authors say the "instrumentation" process is highly subjective, but the statement "While neither model produces an instrumentation identical to that of the original, [...] versions are reasonable rearrangements of the song" is hard to backup with an user experiment with ratings for the different instrumentation alternatives.

The work described here could be thought as two subsequent tasks: 1. Separate different voices; 2. Assign an instrument to each of the voices. The 2nd part is quite subjective as stated before, but 1 could be studied as "how well the model learns to separate the different voices" which doesn't need a user test and will get to know more about the results.

At the end of the discussion section it is said that this could be used as a sort of pre-training for other tasks. While I also think so, I think it would be better if the authors give a few examples on that.

The work presented looks very interesting for creative purposes, and definitively can help musicians to create sort of "new coherent instrumentations" for symbolic scores.

View Meta-Reviews

Paper ID
293

Paper Title
Towards Automatic Instrumentation by Learning to Separate Parts in Multitrack Music

Track Name
Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. The title and abstract reflect the content of the paper.
Disagree

3. The paper discusses, cites and compares with all relevant related work.
Disagree

4. The writing and language are clear and structured in a logical manner.
Agree

5. The paper adheres to ISMIR 2021 submission guidelines (uses the ISMIR 2021 template, has at most 6 pages of technical content followed by "n" pages of references, references are well formatted). If you selected "No", please explain the issue in your comments.
Yes

6. The topic of the paper is relevant to the ISMIR community.
Agree

7. The content is scientifically correct.
Agree

8. The paper provides novel methods, findings or results.
Disagree

9. The paper provides all the necessary details or material to reproduce the results described in the paper.
Disagree

10. The paper provides reusable insights (i.e. the capacity to gain an accurate and deep understanding). Such insights may go beyond the scope of the paper, domain or application, in order to build up consistent knowledge across the MIR community.
Disagree

11. Please explain your assessment of reusable insights in the paper.
The paper proposes a novel task on keyword adaptation of multi-instrument scores and an approach to address it.

15. The paper will have a large influence/impact on the future of the ISMIR community.
Disagree

17. Overall evaluation
Weak Accept

18. (Initial) Main review and comments for the authors
This paper addresses the task of voice assignment in symbolic representations, in the context of adapting multi-instrument pieces to keyword setup. However, the work just focuses on the voice assignment task, proposing an architecture that is evaluated against three baseline approaches. I think the approach is technically sound, the methodology is well designed and the results are relevant contributions to the field. However, I think the paper has some drawbacks, which I think should be addressed before its acceptance or commented as limitations in the paper. Some of these drawbacks are somehow major and do not fit the timeline or procedure of minor reviews

which are currently implemented at ISMIR, so for this I would recommend weak rejection of the paper.

- Title and abstract:
o The title is a bit confusing from some readers as the term “multitrack” is widely used to refer to audio and there is a reference to “automatic instrumentation”, which is not a widely used term. I would suggest the authors to consider the keyword “score” or “symbolic” to communicate it is about symbolic representations.
o The paper has as an ultimate goal the generation of instrumentation for keyword players, but there is no evaluation or adaptation carried out to adapt the output to keyword playing. So I would also suggest to discard the “automatic instrumentation” part in the title.

- Quantitative evaluation: I find the evaluation as relevant but I feel the baseline models are quite simple, so it is somehow expected that a trained model would overperform them. Could the authors consider any SOTA machine learning model as a baseline? One idea would be to check the current state of the art in multipitch estimation as voice assignment is part of this process, and there are some algorithms already proposed.
- Qualitative evaluation: I find that the qualitative results section of the paper is very superficial while I think it is one of the main challenges of generative models: user-centric evaluation. For instance, the authors mentions that the ultimate goal is to generate keyword scores, but there is no requirements of the systems attached to this goal, no evaluation on this context. I think the quantitative evaluation carried out should be complemented by a user perceptual study to reflect on the musical quality of the voicing assignment.

---

19. Meta-review and final comments for authors

After analyzing all reviews and discussing about the paper contributions and limitations in the forum, we all agree that the paper has nice contributions to the MIR field and could fit the ISMIR conference.
