View Reviews

Paper ID
78

Paper Title
Improving Choral Music Separation through Expressive Synthesized Data from Sampled Instruments

Track Name
Papers

Reviewer #1

Questions

2. I am an expert on the topic of the paper.
Strongly agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

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

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

6. Please justify the previous choice (Required if "Strongly Disagree" or "Disagree" is chosen)
While the paper discusses the relevant related work (Sec. 2), I believe the paper would benefit from a comparison to at least one baseline in terms of performance - adding a comparison to a baseline system besides comparing the different proposed scenarios.

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 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

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Agree

10. Scholarly/scientific quality: The content is scientifically correct.
Strongly agree

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.

Agree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Strongly agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: 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.
Strongly agree

16. Please explain your assessment of reusable insights in the paper.
The proposed synthesis pipeline can be used in other types of repertoire when the appropriate sampled instrument libraries are available. Besides, the authors show that pre-training on such synthetic data improves the performance of source separation for tasks where the amount of real data is very limited.

17. Write ONE line (in your own words) with the main take-home message from the paper.
Generation of synthetic data via sampled instruments to pre-train source separation models improves the performance of the models on real data.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Weak accept

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
This paper proposes a pipeline for systematically synthesizing choral music to pretrain deep learning models for source separation (SS). The authors show that this pretraining step is beneficial for the model when fine-tuned and evaluated on real-world choir recordings. Although the techniques are not novel, the approach is especially interesting when combined with the synthesis pipeline.

While I believe this work has some strong points, there are some concerns that the authors would need to address for the paper to be accepted for publication. See the list below:

* The biggest concern I have about the paper is the missing comparison to baseline system(s). As mentioned in Sec2, SS has been applied to choral music in a handful of recent papers. I think it's crucial to add at least the quantitative comparison to one baseline system to position this work within the existing methods (e.g. [13], [15], as there's an evaluation data overlap).

* It is not entirely clear to me if the dataset will be released, or only the code for the pipeline will be public. I believe it

is essential for this work to release the dataset together with score-based annotations. Listening to the results, it would be very valuable data to push the state-of-the-art forward.

* Some further explanation about the lyrics/word setting would help understand this part of the pipeline. For the "random words", do you assign a random word to each note/phrase and voice part? do you use the same word for all voices?

* I think it would be good to know how much data is used for training in the "None" pre-training setting (for each dataset, e.g., number of songs or minutes). I would expect this to be very small, especially for CSD, hence overfitting.

Minor comments:

* In line 18-19: you mention other works that evaluation choral music separation, so I would not say this paper is the first to evaluate this task "comprehensively".
* In Table 1: please, check the spelling: "Choral Singing Dataset" (without the "e", also in other places of the paper), and DCS Dataset is actually called "Dagstuhl ChoirSet (DCS)". Also, the number of songs for Bach and Barbershop Collection is 26 in total? In other works they contain 22 and 26 songs, respectively, making the total 48 songs.
* Line 82-83: please mention that U-Net was not originally designed for music-related tasks (biomedical image segmentation).
* The sentence in lines 121-122-123 reads strange: "the beginning of a sizeable dataset" is a very vague statement.
* In lines 136-139 it's worth mentioning that multi-pitch estimation can be applied to the mixture and used as conditioning, as it's not score-based.
* The "octave shifting" process in line 226 is not entirely clear: do you shift the full voice part or just individual notes that are out of the range? If the latter, this produces non-realistic jumps in the resulting melody. Please, clarify the process either way.
* The paragraph between lines 277 and 295 has a lot of technical details about the networks. I would suggest to separate them, as it becomes a bit difficult to read.
* The provided demos are very useful to understand the different synthesis conditions as well as the results. Great idea!

Overall, I think the paper has some nice insights and interesting results. However, the lack of comparison to baselines makes it difficult to assess within the context of choral music separation.

Reviewer #2

Questions

2. I am an expert on the topic of the paper.
Agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the “Special Call” column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly agree

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

6. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen)
There are a few citations that would be good to add, or statements that would be good to support with a citation, if relevant literature exists:

"However, due to the lack of harmonic information in the spectrogram, time-domain models do not typically achieve performance parity with frequency-domain models in the field of musical instrument separation": this seems intuitively right to me, but are there any suitable references for this? I can certainly see harmonic information becoming more important when singing or speaking voice is mixed with music, and this seems more important compared to just separating overlapping speakers.

"but it is more difficult to model the acoustic features of these four voices than piano.": it would be good to add some more detail for this statement, and cite any relevant literature.

Add citation for SDR, I think this one is appropriate:
Vincent, Emmanuel, Rémi Gribonval, and Cédric Févotte. "Performance measurement in blind audio source separation." IEEE transactions on audio, speech, and language processing 14.4 (2006): 1462-1469.
https://hal.inria.fr/inria-00544230/document

In second paragraph of section 4.3, include citations in the text for the datasets described (these citations are in Table 1, but it would also be good to have them in the text)

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Strongly agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 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

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Agree

11. Please justify the previous choice (Required if “Strongly Disagree” or “Disagree” is chosen)
I see one minor issue with the experiments: for the real data, the paper experiments with different splits into train and evaluation data. However, it seems that by doing this, the examples in the evaluation set change for each split, i.e. the evaluation set for 10% train / 90% eval is larger and different than the evaluation set for 40% train / 60% eval. Thus, the evaluation scores across different train/eval splits are not apples-to-apples. It would be better if the author used the same fixed evaluation set across all splits, say the smallest evaluation set for the 70% train / 30% eval split. I don't expect this would change the results or conclusions of the paper.

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Agree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the

scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: 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

16. Please explain your assessment of reusable insights in the paper.
This is an instance of constructing synthetic training data to help MIR systems generalize to real-world data, and I think that such instances can be quite useful. E.g. perhaps the expressive parameters used to synthesize voice could be extended to separating vocals from accompaniment for more general types of music. However, the caveat is that it may take some effort to tune the synthesizer parameters appropriately.

17. Write ONE line (in your own words) with the main take-home message from the paper.
Pre-training a model with a larger amount of well-crafted synthetic data and fine-tuning on a small amount of real data improves choral music separation on real data.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Weak accept

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
Summary of the paper:
This paper addresses the task of separating mixtures of singers for choral music. This particular domain suffers from a lack of training data, so this paper proposes to construct synthetic training examples using MIDI and well-crafted virtual instrument plugins. The paper tries multiple models (Spec U-Net, Res-U-Net, Wave-U-Net, Conv-TasNet) for this particular task, and shows for the best model (Spec U-Net) that pre-training on well-crafted synthetic data, then fine-tuning on a small amount of real data, achieves the best results on real data for choral music separation, in terms of SDR. The paper also includes an experiment on the URMP dataset (results in supplementary).

I would have given a higher score if some of the issues I raise in my "other comments" below were more clear.

Strengths:
1) The paper is well-written and well-organized.

2) The results are convincing, as they are evaluated on multiple datasets, which show that the proposed approach is effective.

3) Detailed audio demos and code are provided.

Weaknesses:
1) A few details are a little unclear (see my other comments below)

2) This method is perhaps not very generally applicable, in that good parameters for the synthesizers have to be selected for any particular scenario. This means that tuning the synthetic data for a new scenario may take some work.

Other comments:
1) The paper produces 8.2-hour-long dataset for 347 pieces at 90 BPM. Would there be any advantage to generating more data at different BPM?

2) "Deep learning methods for audio source separation are gradually outperforming traditional methods (e.g. Non-negative Matrix Factorization [5]). Separation models have gradually been developed following two directions: frequency-domain models and time-domain separation models.": the word "gradual" is repeated. Also, for general audio source separation, I think the deep networks have been outperforming NMF for quite some time, so there is nothing gradual about it. I think it would be good to revise this discussion. For papers that specifically compare deep network-based methods to NMF, here are a few of the earlier ones (2014-2017) I could find with a quick search:
https://sapl.gist.ac.kr/wp-content/uploads/2017/01/NMF-based-target-source-separation-using-deep-neural-network.pdf
https://arxiv.org/pdf/1508.04306.pdf
https://www.merl.com/publications/docs/TR2015-029.pdf
https://arxiv.org/pdf/1709.07124.pdf

3) A suggestion for the audio demos: especially since the audio files are long, displaying spectrograms alongside the audio widgets would allow quicker browsing of the demos.

4) Generally, I think it is best to use just one decimal place for units in decibels, since humans can often not even tell the difference within 0.1 dB.

5) What is the audio sampling rate used for experiments? I might have missed it, but wasn't able to find it.

6) "Time-domain models can model the piano acoustic features well to achieve a good performance, but find it hard to model the vocal features solely on the waveform and face the drops in the vocal dataset.": what are the "drops"? Does this mean the gaps between syllables in singing voice, or the silences before singers start or when they stop? Also, I don't see what this has to do with time-domain models.

7) Why did the paper train on synthetic, then fine-tune on real? How would training simultaneously on synthetic and real compare to this?

8) "For evaluation, source-to-distortion ratio (SDR) is one of the most widely used metrics for evaluating a source separation system’s output, which measures a ratio between the original source track and the noise, interference, added artifacts in the separation track.": this is missing the part of SDR that estimates a 512-tap FIR filter applied to the reference signal. I understand that this metric is often used for music separation, but is this invariance justified? This paper (https://arxiv.org/pdf/1811.02508.pdf) discusses some of the potential pitfalls, which can be especially bad when the separation model over-suppresses signals. From the demos, I don't hear much over-suppression, so SDR is probably fine here, but this is something the authors should probably be aware of, and perhaps briefly discuss in the paper.

Typos and writing improvements
1) "receives relative less" -> "receives relatively less"

2) Use present tense when describing the network: "The masked latent features will be decoded" -> "The masked latent features are decoded"

3) "Res-U-Net achieves very closely.": doesn't parse. How about "The Res-U-Net achieves very close performance"

4) "with the early stop in a 10-epoch patience" -> "with early stopping using a 10-epoch patience"

5) "Data augmentations as ``octave shifting``" -> "Data augmentations of ``octave shifting``"

6) "Thus, pertaining on" -> "Thus, pretraining on"

Reviewer #3

Questions

2. I am an expert on the topic of the paper.
Agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly agree

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

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Strongly agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 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

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Agree

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of

existing methods from other research fields are considered novel at ISMIR conferences.
Agree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: 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.
Strongly agree

16. Please explain your assessment of reusable insights in the paper.
Recently, the quality of virtual instruments has increased sufficiently, and they are replacing real instruments in various music. At this point, the supplementary material would be used as a starting point for generating sufficient amount of high-quality instrument database for many separation/transcription tasks.

17. Write ONE line (in your own words) with the main take-home message from the paper.
A VSTi-based midi-to-wave pipeline and its source code is proposed.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Disagree

21. Overall evaluation: Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines
Weak accept

22. Main review and comments for the authors. Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
This paper proposes a pipeline for synthesizing a high-quality choral/instrument database, which is very important bottleneck in various separation tasks suffering from lack of a large amount of high-quality database. The paper is well organized and contains sufficient experiments to support the proposal, therefore, it is confirmed that the synthesized choral music database robustly improves the voice separation performance of real choral music. Although, the proposed pipeline is not a completely new approach, supplementary material would be the first public example of generating high-quality audio from a midi file using a virtual/sampled instrument and its control parameters. The material is expected to be widely used in various separation and transcription tasks where it is difficult to obtain a high-quality database.

However, it is not easy to synthesize an audio with only the provided supplementary material, which is the reason for the weak accept. I would recommend to adding detailed example of synthesizing a choral or piano using a free virtual/sampled instrument and synthesis configuration for realistic sound. It would be helpful for many researchers unfamiliar with DAW or VSTi.
There are concerns that the hyperparameters of the separation network used in the experiment are not optimized

values and the time and database may be insufficient to train the network. If the pretrained network of speech separation was used, there would be no significant difference in the results, but the model selection process would have been simpler.

View Meta-Reviews

Paper ID
78

Paper Title
Improving Choral Music Separation through Expressive Synthesized Data from Sampled Instruments

Track Name
Papers

META-REVIEWER #1

META-REVIEW QUESTIONS

2. I am an expert on the topic of the paper.
Agree

3. Does this submission relate to the topics mentioned in the Special Call for Papers on Cultural and Social Diversity in MIR? Please refer to the Call For Papers - Special Call section regarding the scope of this special call. Please also take into account the intention of the authors by checking the "Special Call" column in the Reviewer Console.
No

4. The title and abstract reflect the content of the paper.
Strongly disagree

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

7. Readability and paper organization: The writing and language are clear and structured in a logical manner.
Agree

8. The paper adheres to ISMIR 2022 submission guidelines (uses the ISMIR 2022 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

9. Relevance of the topic to ISMIR: The topic of the paper is relevant to the ISMIR community. Note that submissions of novel music-related topics, tasks, and applications are highly encouraged. If you think that the paper has merit but does not exactly match the topics of ISMIR, please do not simply reject the paper but instead communicate this to the Program Committee Chairs. Please do not penalize the paper when the proposed method can also be applied to non-music domains if it is shown to be useful in music domains.
Strongly agree

10. Scholarly/scientific quality: The content is scientifically correct.
Agree

12. Novelty of the paper: The paper provides novel methods, applications, findings or results. Please do not narrowly view "novelty" as only new methods or theories. Papers proposing novel musical applications of existing methods from other research fields are considered novel at ISMIR conferences.
Agree

13. The paper provides all the necessary details or material to reproduce the results described in the paper. Keep in mind that ISMIR respects the diversity of academic disciplines, backgrounds, and approaches. Although ISMIR has a tradition of publishing open datasets and open-source projects to enhance the scientific

reproducibility, ISMIR accepts submissions using proprietary datasets and implementations that are not sharable. Please do not simply reject the paper when proprietary datasets or implementations are used.
Agree

14. Pioneering proposals: This paper proposes a novel topic, task or application. Since this is intended to encourage brave new ideas and challenges, papers rated "Strongly Agree" and "Agree" can be highlighted, but please do not penalize papers rated "Disagree" or "Strongly Disagree". Keep in mind that it is often difficult to provide baseline comparisons for novel topics, tasks, or applications. If you think that the novelty is high but the evaluation is weak, please do not simply reject the paper but carefully assess the value of the paper for the community.
Disagree (Standard topic, task, or application)

15. Reusable insights: 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

16. Please explain your assessment of reusable insights in the paper.
The paper shows that when using synthetic data for training a choral separation system, making sure the data is as realistic as possible will improve the performance of the trained system.

17. Write ONE line (in your own words) with the main take-home message from the paper.
Synthesizing realistic vocals with expressivity can help train a choral music separation system.

20. Potential to generate discourse: The paper will generate discourse at the ISMIR conference or have a large influence/impact on the future of the ISMIR community.
Agree

21. Overall evaluation (to be completed before the discussion phase): Please first evaluate before the discussion phase. Keep in mind that minor flaws can be corrected, and should not be a reason to reject a paper. Please familiarize yourself with the reviewer guidelines at https://ismir.net/reviewer-guidelines.
Weak accept

22. Main review and comments for the authors (to be completed before the discussion phase). Please summarize strengths and weaknesses of the paper. It is essential that you justify the reason for the overall evaluation score in detail. Keep in mind that belittling or sarcastic comments are not appropriate.
The paper addresses the problem of vocal polyphony source separation which is a challenging and understudied problem. It provides a method for synthesizing a realistic training dataset and assesses the performance gain of using the generated data as a training set with a recent separation method. The paper shows that the generated data makes it possible to improve separation performance, especially in a data scarcity scenario.

The paper is overall interesting and quite easy to understand. There are some paragraphs that are hard to read and would benefit from rewriting though (see further).
The evaluation part is quite convincing and brings interesting insights on the use of realistic synthetic data for the studied task.
The overall contribution is rather modest (the vocal synthesis pipeline is mainly basic engineering on top commercial pieces of software, and the experiment consists in training existing systems on several kinds of generated data) but nonetheless very relevant to the ISMIR community, and there is a significant effort for reproducibility as the code of the experiments was provided with the supplementary materials. There is no explicit mention of releasing the dataset, but I think it would be a significant addition to the overall contribution.

The supplementary material provides examples of both sound examples of the dataset and of the output of the separation systems. It is a nice addition that helps understand the contribution of the paper.

Some comments to be addressed:

“time-domain models do not typically achieve performance parity with frequency-domain models in the field of musical instrument separation” => well this is not totally true and the best models to date can use (see for instance the results of the Music Demixing Challenge @ISMIR 2021) hybrid architectures that leverage both time-domain and frequency-domain inputs (eg Demucs v3).

I’m not sure I get why there are three scenarios (3 different ratios) in the experiment reporter in Table 4 / figure 2. The lowest ratio seems to correspond to a few shot settings and the highest to a fully supervised fine-tuning, but the intermediary scenario does not really make sense, is not really commented on, and does not bring much to the result. I would recommend removing it or explaining further the insights we could get from the results for this intermediate scenario.

“Deep learning methods can improve this expressiveness modeling.” This is too vague and does not carry much information. Some references or some more specific explanations should be provided to support this claim.

typo/reformulation:
l8: “of of synthesized training data” -> “of synthesized training data”
l18: “To best of our knowledge” -> “To the best of our knowledge”
l21: “qualify” -> “quality”.
“Audio source separation is a basic information retrieval task” => I would not qualify source separation as an information retrieval task. An information retrieval task involves retrieving information from a collection of resources. I think the confusion could come from “Music Information Retrieval” which encompasses tasks that can hardly be qualified as an information retrieval task, still, I would definitely recommend changing the formulation. “inverse problem” or “music signal processing task” would be much more adapted, IMHO. Also qualifying a task as “basic” may sound like it is easy, so I would avoid this adjective as well.
“Research in choral music separation receives relative less attention.” -> relative(ly) to what?
“However, due to the lack of harmonic information in the spectrogram, time-domain models do not typically achieve performance parity with frequency-domain models in the field of musical instrument separation.” => this sounds very strange. I don’t see why there is a lack of harmonic information in the spectrogram and why it would cause the time-domain models to have lower performance.
“We continue to demand unconditioned choral music separation” -> I don’t get the message in this sentence (and the whole paragraph around is a bit obscure as well). It needs reformulation.
“Human singing voice is also considered as one of sampled instrument types.” => “The human singing voice is also considered as an instrument type”?
“the large difference” -> “the largest difference”

---

23. Final recommendation (to be completed after the discussion phase) Please give a final recommendation after the discussion phase. In the final recommendation, please do not simply average the scores of the reviewers. Note that the number of recommendation options for reviewers is different from the number of options here. We encourage you to take a stand, and preferably avoid “weak accepts” or “weak rejects” if possible.
Accept

---

24. Meta-review and final comments for authors (to be completed after the discussion phase)
The reviewers agreed that the paper could be accepted for publication.

They notably mentioned that:
- The topic of the paper is interesting, novel, relevant to ISMIR and scarcely addressed in the literature.
- The evaluation is convincing and shows that the use of realistic synthetic data can improve system performances.

- The paper is well-written and well-structured.

Though, there are issues that the reviewers strongly encourage the authors to address:
1) the artificial vocal dataset should be released.
2) The evaluation lacks an external reference to compare to. Notably, the authors should add a performance on BBC from [13] and [15] to compare with their proposed method.