View Reviews

Paper ID	88
Paper Title	CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models
Track Name	WASPAA2023-Main

Reviewer #1

Questions

1. How confident are you in your evaluation of this paper?
3. Very Confident

2. Importance/Relevance
3. Of sufficient interest

4. Novelty/Originality
4. Very original

6. Technical Correctness
3. Probably correct

8. Experimental Validation
4. Sufficient validation / theoretical paper

10. Clarity of Presentation
3. Clear enough

12. Reference to Prior Work
4. Excellent references

14. Overall evaluation of this paper
4. Definite accept

15. Detailed assessment of the paper
The authors propose a method for text-to-audio synthesis without text-audio pairs by using the diffusion model and the pretrained CLIP model. The proposed method is a simple and powerful way to solve the problem of small text-audio pair data. Evaluation experiments have also shown the effectiveness of the proposed method. The effectiveness of the proposed method is appropriately demonstrated in the evaluation experiments. The reviewer is confident that this paper is of sufficient quality to be accepted.

It would be better to include details of the proposed method and network settings, but it is quite understandable that there is a limit to the paper space.

Reviewer #2

Questions

1. How confident are you in your evaluation of this paper?
2. Confident

2. Importance/Relevance
3. Of sufficient interest

4. Novelty/Originality

4. Novelty/Originality
3. Moderately original

6. Technical Correctness
3. Probably correct

8. Experimental Validation
3. Limited but convincing

10. Clarity of Presentation
3. Clear enough

12. Reference to Prior Work
3. References adequate

14. Overall evaluation of this paper
4. Definite accept

15. Detailed assessment of the paper
The authors present a system for generating (environmental) sounds from text prompts. The system is trained without text-audio pairs. Instead the authors leverage a database of unlabled videos (including audio track) and existing text-image embeddings. Results stay below systems leveraging text-audio pairs but still seem promising and could improve even further in the future, as image/video datasets tend to be several orders of magnitude bigger than audio datasets.

Overall, a solid contribution from my perspective. While the approach might be limited in utility for specific domains (e.g. music) it certainly looks very promising for general sound synthesis and environmental sound generation in particular. Experimentation looks proper and the ablation against the guidance parameter is interesting. Further, the design (in particular the inclusion of a prior to overcome a domain gap as described in the paper) shows that the issues faced during the earlier stages of the system were identified and understood, and correct conclusions were drawn.

And while the results indicate that using text-audio pairs still yields better systems, this is certainly important work that is valuable across a wider range of use cases.

Reviewer #3

Questions

1. How confident are you in your evaluation of this paper?
2. Confident

2. Importance/Relevance
3. Of sufficient interest

4. Novelty/Originality
3. Moderately original

6. Technical Correctness
4. Definitely correct

8. Experimental Validation
3. Limited but convincing

10. Clarity of Presentation
3. Clear enough

12. Reference to Prior Work
3. References adequate

14. Overall evaluation of this paper
3. Marginal accept

15. Detailed assessment of the paper

Summary

This paper leverages the CLIP model to cross the gap between image and text, then train a latent diffusion model to generate audio spectrogram from the image embedding from the CLIP model, thus bypass using the text-audio pairs during training. To mitigate the modality mismatch between text and image, the authors propose to train a prior diffusion to link the embedding and image embedding from CLIP, which further improve the quality of generated audio given the text input.

Major, Limited application

- CLIP model is trained for a single image and the related text description. So in this work, if I understand well, the author use only one image as the input to extract the visual embedding as input. This makes sense for relatively static video (e.g. a train is coming from afar), but I'm wondering if this is applicable to more dynamic scenes or longer sequences.
- Moreover, the text description should be OK to generate some simple voice or ambient sound, but I don't think this contains enough information for semantic generation, such as music or human speech. Given the using of pretrained CLIP model, this method may be hard to extend to these directions

Minors

- In Figure 2, the description for the latent diffusion is somehow ambiguous, it looks like the model is only trained with on step of diffusion (i.e. from x_t to x_{t-1})

View Meta-Reviews

Paper ID	88
Paper Title	CLIPSonic: Text-to-Audio Synthesis with Unlabeled Videos and Pretrained Language-Vision Models
Track Name	WASPAA2023-Main

META-REVIEWER #2

META-REVIEW QUESTIONS

1. Please provide your meta-review based on the summary of the reviews. Please encourage the discussion among reviewers to arrive at a consensus. While your meta-review is an organic aggregation of the reviews, please also provide your own opinion and recommendations. Your meta-review is visible to the authors, so please be constructive. In this section, please provide your summary comment. We will ask for your verdict in a separate question.

The paper introduces a novel method for text-to-audio synthesis that overcomes the reliance on text-audio pairs by leveraging the diffusion model and the pretrained CLIP model. The proposed approach demonstrates simplicity and effectiveness in addressing the challenge of limited text-audio pair data. Evaluation experiments validate the effectiveness of the method, showcasing its potential for generating audio from text prompts.

Reviewers appreciate the thoroughness of the evaluation and the demonstrated efficacy of the proposed method. The inclusion of the ablation study on the guidance parameter and the utilization of a prior diffusion to bridge the domain gap reveal a comprehensive understanding of the system's design. This indicates that the authors have identified and addressed previous issues encountered during system development, drawing accurate conclusions.

While the results show that directly utilizing text-audio pairs still yields superior systems, reviewers recognize the value and broader applicability of the proposed method. It is deemed promising for general sound synthesis and particularly suitable for generating environmental sounds. The experimentation process is considered appropriate, and the analysis of the latent diffusion in Figure 2 is appreciated, despite some ambiguity in its description.

However, there are concerns raised by reviewers regarding the method's limited application in more dynamic scenes or longer sequences, as the reliance on a single image for extracting visual embeddings from the CLIP model might not be suitable in those scenarios. Additionally, reviewers question the method's capability to generate more semantically complex audio, such as music or human speech, due to the constraints imposed by the pretrained CLIP model. Extending the method in these directions may pose challenges.

To further enhance the paper, reviewers suggest providing additional details about the proposed method and network settings, which could improve understanding and reproducibility. Despite these minor concerns, the paper is considered of sufficient quality and value, with potential for future improvements and extensions.