AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Haohe Liu*, Zehua Chen*, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, Mark D. Plumley
• Many thanks to other co-authors who made this work possible:
  • Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, Mark D. Plumbley
What is Audio Generation

Definition, history, and related works
Audio Generation

• The creation of sound through various ways

• The targets include:
  • *Sound Effect* (Natural, Human-made objects, Animal, etc.)
  • *Speech* (Emotion, Pace, Gender, etc.)
  • *Music* (Genre, Rhythm, Instruments, etc.)
  • *Other* (Imaginary sound, compositional sound)
History of Sound Effect Creation

Foley Artist
Recreation of the realistic ambient sounds
Jack Foley
Modern foley art

Physical Modeling
Synthesis by modeling physical process
Generate sound based on shape, material, strength, and excitations.

Sound Effect Library
Digital collection of sound effect
Sound Ideas
BBC SFX
Freesound...

Add live sound effects 1920s...
Modern Foley Artist

3/6/23
History of Speech Creation

**Machanical Synthesis**

- Simulating vocal tract, tongue, and lips
- *Kratzenstein Resonators*
- Kempelen’s Speaking Machine

**Electronic Signal Processing**

- Synthesis by modeling physical process
  - The VODER
  - Concatenation synthesis
  - Formant synthesis
  - Articulatory synthesis

**Deep learning-based**

- Digital collection of sound effect
  - Tacotron
  - FastSpeech
  - NaturalSpeech
  - VALL-E

---

Christian Gottlieb Kratzenstein (1723-1795)
Kratzenstein’s resonators that can produce: [aː], [eː], [iː], [oː] and [uː]

Kempelen’s speaking machine (replica, 1837)

The Vocal Demonstrator (VODER, 1939)
History of Music Creation

Musical Instrument

- a device created or adapted to make musical sound
- Aurignacian flute
- Kempelen’s Speaking Machine
- ... 

Synthesizers, MIDI, and DAWs

- Create music with electrical devices
  - Moog Synthesizer
  - Modern DAW
  - MIDI
  - ...

- Create music with machine learning
  - Synthesizer (DDSP)
  - Symbolic / MIDI (MuseNet)
  - Waveform (JukeBox)
  - Controllability (MIDI-DDSP)
  - ...

Deep learning-based

Modern Musical Instruments

- Aurignacian flute
  (43000 and 35000 years ago)

The Moog Synthesizer by Robert Moog (1970s)

Digital Audio Workstation (DAW)

MuseNet by OpenAI (2019)

DDSP by Google (Engel et al., 2020)

MIDI-DDSP by Yusong et al. (2021)

JukeBox by OpenAI (Dhariwal, 2020)
Can machine do general audio generation?

- I’m a
  - foley artist,
  - musical instruments performer,
  - oral broadcaster,
  - sound imaginer,
  - ...

- Communicate with AI by natural language
  - Text-to-Audio Generation
Why: Text-to-Audio Generation

Applications, and motivations
Text-to-Audio Generation Usage Cases

• Computational “foley artist”: (e.g., https://www.thefoleybarn.com)
  • Game developer: e.g., A ghost is haunting a house.
  • Audio producer: e.g., high heels hitting metal ground.
  • Movie producer: e.g., the laser sound from a laser gun.
  • ...

• Automatic content creation (> 60 startups)
  • Endless music
  • Audiobook with ambient noises
  • White noise for meditation
  • ...

• Data Augmentations

1https://github.com/csteinmetz1/ai-audio-startups

Sound is often the unsung hero of the movie world
- Hans zimmer
Text-to-Audio Generation Usage Cases

• Text is a bridge between audio and other modalities

Audio Captioning

Text-to-Audio Generation

Text

Audio

Other Modalities

e.g. Image, Video, Sketch

1https://github.com/csteinmetz1/ai-audio-startups
Generation VS Retrieval

**Efficiency**
- No need for retrieval
- Endless audio samples
- Fine-grained control on sound
  - Emotion, pitch, materials, etc.
- Future way of fuzzy data storage
  - 2GB VS 2048 GB

**Creativity**
- Generate non-exist sound
  - e.g., Half cat Half sheep sound
- Inspire the content creation
Related works

Introduction, and comparison
Related works

• Label-to-Audio Generation
  • Acoustic Scene (Kong et al., 2019), Sound event (Liu et al., 2019), FootStep (Comunit et al. 2019), ...

• Text-to-Audio Generation
  • DiffSound (Yang et al., 2022), AudioGen (Kreuk et al., 2022), Make-an-Audio (Huang et al., 2023)

• Text-to-Music Generation
  • MusicLM (Andrea et al., 2023)
  • Moûsai (Flavio et al., 2023)
  • Noise2Music (Huang et al., 2023)

• Others
  • JukeBox (Dhariwal et al., 2020), AudioLM (Borsos et al., 2022), SingSong (Donahue et al., 2023),...
DiffSound (Yang et al., 2022)
AudioGen (Kreuk et al., 2022)
Make-an-Audio (Huang et al., 2023)
Related works

DiffSound (Yang et al., 2022)

AudioGen (Kreuk et al., 2022),

Make-an-Audio (Huang et al., 2023)
Comparison with previous studies

• Previous audio generation studies:
  • Requires large-scale audio-text pairs
    • Prev: Text → Audio → Loss → Backprop
    • Our: Audio → Audio → Loss → Backprop
  • High computational cost
    • Prev: 64 or 32 V100 GPUs (AudioGen, DiffSound)
    • Our: 1 GPUs
  • Limited generation quality and diversity.
  • Discrete latent space may limit model performance

Previous works: 10+ datasets, 800K audio-text pairs (still not enough).
Self-supervised Learning for Audio Generation!
Self-supervised Audio Generation

**Human Developer:**
- Here are some audio-text pair, try to figure out their relation!

**Step 1**
- Audio → Text

**Step 2**
- Audio

**Human Developer:**
- Here are more audio data, try to figure out how to generate them using your knowledge!

**AI**
- Nobody knows audio better than me!!!
- Tell me what you want!

**Latent Space**
- Laughing
- Giggling
- Guitar
- Ukelele

[Diagram showing the relationship between audio and text with Latent Space, and text areas for OK Got it and Nobody knows audio better than me!!!]
How: AudioLDM

Methodology, Advantages, Experiment, and Result
1. **Contrastive Language-Audio Learning (CLAP) Encoders**
   - Align audio and text in one space.

2. **Latent Diffusion Models**
   - Learn to generate VAE latent conditioned on CLAP embedding

3. **Mel-spectrogram Autoencoder**
   - Learn latent representations.

4. **Mel-to-Waveform Vocoder**
   - Reverse Mel back to waveform
Step 1: Contrastive Language-audio Pretraining

Contrastive Language-Audio Pretraining (Wu et al., 2022)

Audio Waveforms

Text Data

- Sentences
- Labels
- Keyword-to-Sentence Augmentation

Text Encoder

Audio Encoder

Laughing is similar to Giggle (Text).
Laughing also sounds like Giggle (Audio).
Laughing is \textbf{not} similar to Guitar (Text).
Laughing does \textbf{not} sound like Guitar (Audio).

New data

OK Got it

Latent Space

Audio

Text

Laughing

Giggling

Guitar

Ukelele

Text

Audio
Step2: Self-supervised Audio Generation Training

Human Developer:
Here are more audio data,
Try to figure out how to generate them using your knowledge!

Nobody knows audio better than me!!!
Tell me what you want!
Advantages of self-supervised training

• Scale up training data easily!
  • Collect Audio → Train model!
• Perform data augmentation easily!
  • Previous works:
    • Mixup (Kreuk et al., 2022)
      • Text1 + Text2 → Audio1+Audio2
    • Pseudo prompt enhancement (Huang et al., 2023)

Make-an-Audio (Huang et al., 2023)
Overall Advantages

• **Less computation cost**
  • Latent Diffusion Models.

• **Less dependency on audio-text pairs.**
  • Train LDMs by self supervision

• **Continuous latent space**
  • Zero-shot audio style transfer.
  • Zero-shot audio super-resolution
  • Zero-shot audio inpainting.
  • …
Zero-shot down stream tasks

- Audio style transfers
  - Corrupt -> Reverse Diffusion

- Audio inpainting
  - Provide temporal hint during sampling.

- Audio super-resolutions
  - Provide frequency hint during sampling.
Training Data (16 kHz)

- AudioSet
- AudioCaps
- FreeSound
- BBC Sound Effect Library

Finally: 3,302,553 ten-seconds (9000+ hours) audio samples without text labels.

Largest scale so far
Evaluation Metrics

• Subjective evaluation
  • **OVL**: Overall quality
  • **REL**: relevance to text

• Objective evaluation
  • **FD**: Frechet Distance
  • **IS**: Inception Score
  • **KL**: Kullback-Leibler Divergence

<table>
<thead>
<tr>
<th>File name</th>
<th>Text description</th>
<th>Overall impression (1-100)</th>
<th>Relation to the text description (1-100)</th>
</tr>
</thead>
<tbody>
<tr>
<td>random_name.108029.wav</td>
<td>A man talking followed by lights scraping on a wooden surface</td>
<td>80</td>
<td>90</td>
</tr>
<tr>
<td>random_name.108436.wav</td>
<td>Bicycle Music Skateboard Vehicle</td>
<td>70</td>
<td>80</td>
</tr>
<tr>
<td>random_name.116883.wav</td>
<td>A power tool drilling as rock music plays</td>
<td>90</td>
<td>95</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Example questionnaire for human evaluation. The participant will need to fill in the last two columns.
## Result – SOTA comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>Datasets</th>
<th>Text</th>
<th>Params</th>
<th>FD ↓</th>
<th>IS ↑</th>
<th>KL ↓</th>
<th>FAD ↓</th>
<th>OVL ↑</th>
<th>REL ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth</td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>83.61</td>
<td>80.11</td>
</tr>
<tr>
<td>DiffSound† (Yang et al., 2022)</td>
<td>AS+AC</td>
<td>✓</td>
<td>400M</td>
<td>47.68</td>
<td>4.01</td>
<td>2.52</td>
<td>7.75</td>
<td>45.00</td>
<td>43.83</td>
</tr>
<tr>
<td>AudioGen† (Kreu et al., 2022)</td>
<td>AS+AC+8 others</td>
<td>✓</td>
<td>285M</td>
<td>-</td>
<td>-</td>
<td>2.09</td>
<td>3.13</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AudioLDM-S</td>
<td>AC</td>
<td>✗</td>
<td>181M</td>
<td>29.48</td>
<td>6.90</td>
<td>1.97</td>
<td>2.43</td>
<td>63.41</td>
<td>64.83</td>
</tr>
<tr>
<td>AudioLDM-L</td>
<td>AC</td>
<td>✗</td>
<td>739M</td>
<td>27.12</td>
<td>7.51</td>
<td>1.86</td>
<td>2.08</td>
<td>64.30</td>
<td>64.72</td>
</tr>
<tr>
<td>AudioLDM-L-Full</td>
<td>AS+AC+2 others</td>
<td>✗</td>
<td>739M</td>
<td><strong>23.31</strong></td>
<td><strong>8.13</strong></td>
<td><strong>1.59</strong></td>
<td><strong>1.96</strong></td>
<td><strong>65.91</strong></td>
<td><strong>65.97</strong></td>
</tr>
</tbody>
</table>

Trained on a single 3090 or A100 GPU!
Result – self-supervised LDMs training

• Training with audio can even outperform training with audio-text pairs.

• Reason:
  • **Audio representation is better than Text**
    1. Text labeling sometimes have weak relations to audio
    • e.g., Boats: Battleships-5.25 conveyor space
    2. Text labeling is error-prone
    • Missing labels in text.
    • Text is difficult to include every details.

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
<th>Audio</th>
<th>FD ↓</th>
<th>IS ↑</th>
<th>KL ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>AudioLDM-S</td>
<td>✓</td>
<td>✓</td>
<td>31.26</td>
<td>6.35</td>
<td>2.01</td>
</tr>
<tr>
<td>AudioLDM-S</td>
<td>✗</td>
<td>✓</td>
<td>29.48</td>
<td>6.90</td>
<td>1.97</td>
</tr>
</tbody>
</table>
Result – Super-resolution and Inpainting

• Super-resolution
  • VCTK (Speech)
  • AudioCaps (General Audio)

• Inpainting
  • AudioCaps

<table>
<thead>
<tr>
<th>Task</th>
<th>Super-resolution</th>
<th>Inpainting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>AudioCaps</td>
<td>VCTK</td>
</tr>
<tr>
<td>Unprocessed</td>
<td>2.76</td>
<td>2.15</td>
</tr>
<tr>
<td>Kuleshov et al. (2017)</td>
<td>-</td>
<td>1.32</td>
</tr>
<tr>
<td>Liu et al. (2022a)</td>
<td>-</td>
<td>0.78</td>
</tr>
<tr>
<td>AudioLDM-S</td>
<td>1.59</td>
<td>1.12</td>
</tr>
<tr>
<td>AudioLDM-L</td>
<td><strong>1.43</strong></td>
<td>0.98</td>
</tr>
</tbody>
</table>

Super-resolution: Log-spectral distance
Inpainting: Frechet audio distance
Inpainting

• Examples
  • Use matched text
  • Use un-matched text
Inpainting

- Examples
  - Use matched text
  - Use un-matched text
Result – Other details

- A good CFG scale is around 2.5
  - Large CFG: Less diversity
  - Small CFG: better diversity, less quality
- Different VAE compression levels.
  - 4, 8, 16
- Evaluation on AudioSet
- Sampling Steps (around 100 DDIM).
- Other ablation studies.

Effect of different classifier-free guidance scale
Audio Style Transfer

Drum beats → Ambient Music

Sheep vocalization → Narration, monologue
Audio Style Transfer

Trumpet → Children Singing

Drum beats → Ambient Music

Sheep vocalization → Narration, monologue
More examples

- Audio super-resolution
- Audio inpainting
- Fine-grained generation control:
  - Controls of object materials
  - Controls of acoustic environment
  - Controls of audio pitch
  - Controls of temporal orders
  - ...

3/6/23
More examples

• A stone is hitting a metal plate
• Dance music with strong beats played by multiple instruments
• healthy deep gurgly 10 second burp
• Very windy condition, trying to fly against the wind in a parachute
• A small water steam in a forest with some bird vocalization
• somone slurping noodles long slurp
More examples (wired sound)

- The weirdest sound in existence
- The cry of Cthulhu the terrifying ancestral deity
- A man is speaking backwards creepily and exhaustively
More examples

- Brain bubbles floating in primordial goo
- 漂浮在原始粘液中的脑泡
Interesting resources

• Image-to-Audio
  • https://huggingface.co/spaces/fffiloni/image-to-sound-fx

• AI music album:
  • https://www.latent.store/albums
AudioLDM on Diffuser

Credit to Sanchit Gandhi from Hugging Face

```python
from diffusers import AudioLDMPipeline
import torch import scipy

repo_id = "sanchit-gandhi/audioldm-text-to-audio"

pipe = AudioLDMPipeline.from_pretrained(repo_id, torch_dtype=torch.float16)
pipe = pipe.to("cuda")

prompt = "Techno music with a strong, upbeat tempo and high melodic riffs"

audio = pipe(prompt, num_inference_steps=10, height=512).audios[0]

# save the audio sample as a .wav file
scipy.io.wavfile.write("techno.wav", rate=16000, data=audio)
```
A few take aways here, thanks!

  • AudioLDM: Text-to-Audio Generation with Latent Diffusion Models
• Project Page: https://audioldm.github.io/
• Hugging Face Space:
  • https://huggingface.co/spaces/haoheliu/audioldm-text-to-audio-generation
• Github:
  • Pretrained model: https://github.com/haoheliu/AudioLDM
  • Evaluation tools: https://github.com/haoheliu/audioldm_eval
• Interesting demo website:
  • https://www.latent.store/albums