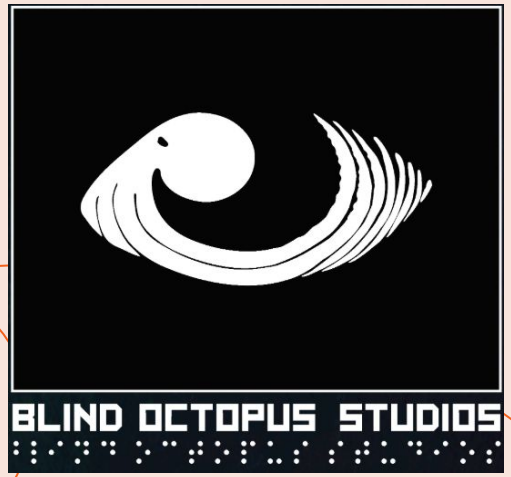
Enhancing Game Accessibility

Sound Generation for Visually Impaired

Norbert Hašan; školiteľ: Lukáš Gajdošech

Motivation



1. Audio is a must have for VI
2. Impact on VI Players
3. Current Production Bottlenecks
4. Current Variation Techniques
5. The Scale of Modern Games
6. Adding AI (neural network)

Thesis Goals

01

Review & Understand

Conduct a comprehensive review of existing sound synthesis methods, from traditional analytical techniques to modern generative neural networks.

02

User-Centric Specification

Consult with the target group of users to specify their unique requirements and preferences for in-game audio cues.

03

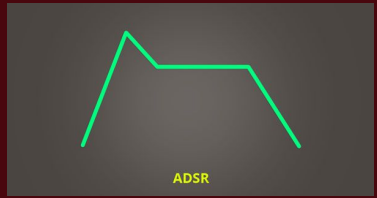
Develop & Implement

Create model for the automatic generation of sound effects based on textual inputs.

Introduction to Sound Synthesis

- What is Sound Synthesis?
  - algorithmically manipulating existing sounds, creating a sound from scratch
- Challenges:
  - Realism & Complexity
  - Variation & Control
  - Computational Cost
- Needs for VI Users:
  - Clarity
  - Informativeness
  - Learnability
  - Non-fatiguing

Sound Synthesis Approaches

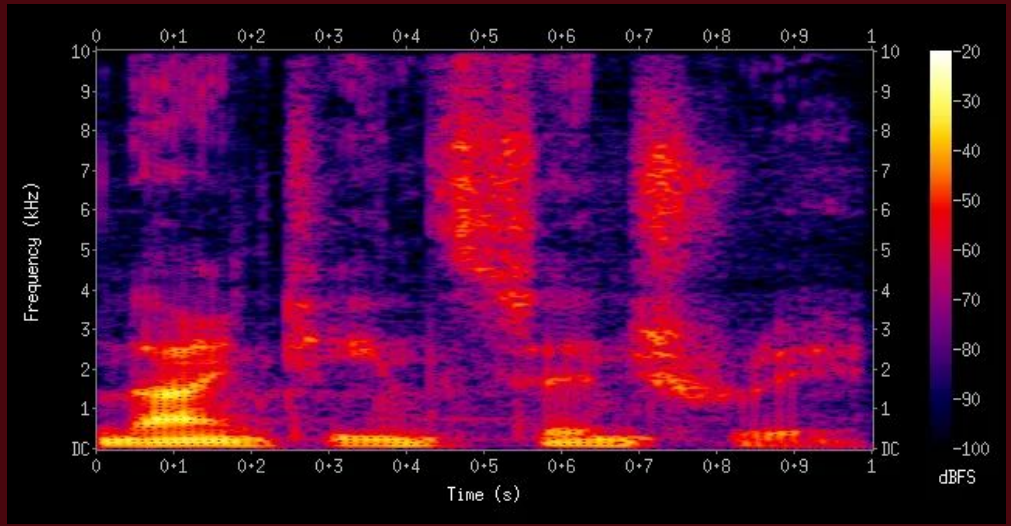


Traditional / Analytical Methods

- **Subtractive Synthesis** — start with a harmonically rich waveform and filter it
- **Additive Synthesis** — combine multiple simple waveforms
- **Frequency Modulation** — modulate the frequency of one oscillator with another
- **Physical Modeling** — simulate the acoustic properties of physical objects/ instruments
- **Procedural Audio** — algorithmically generate sounds in real-time based on game parameters

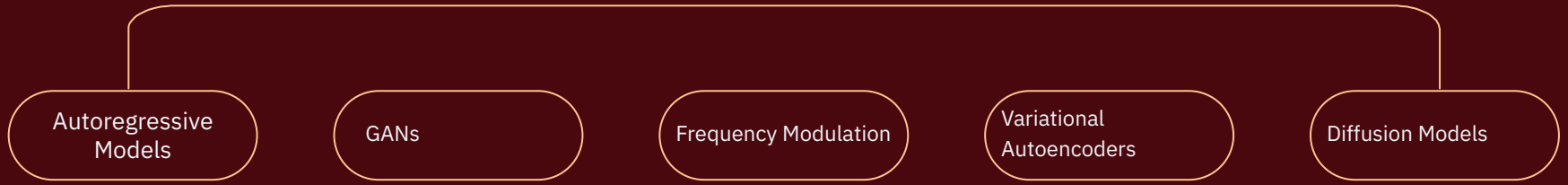
Sound Synthesis Approaches

- Modern Neural Network Approaches:
  - Data-driven: Learn from audio datasets.
  - Captures complex patterns and realism.



Neural Network Sound Synthesis

Model types



generate sample by sample,
+ high fidelity
- slow generation
WaveNet

generator vs. discriminator
+ fast parallel sampling
- local coherence
- unstable training
WaveGAN, GANSynth

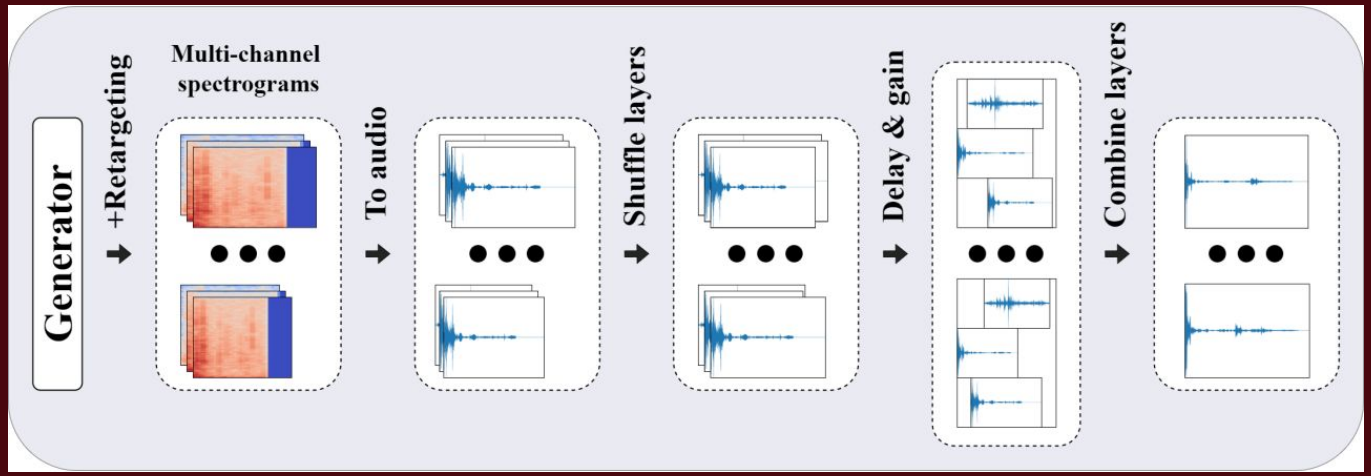
learn invertible transformation
+- balances quality and speed
WaveFlow

Probabilistic latent space.
+ good for interpolation
+control
+ understanding data structure

iteratively add noise to data, reverse it
+generate from small data

Neural Network Sound Synthesis

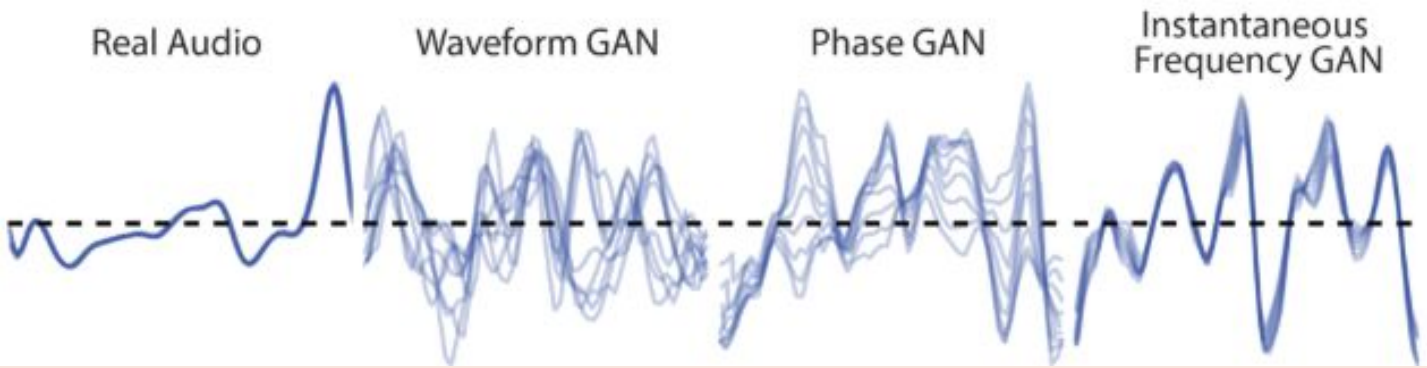
- Raw Waveform:
  - Direct sample modeling.
  - Computationally intensive; potentially highest fidelity.
- Spectrogram-based:
  - Model time-frequency.
  - More compact, easier for models; needs vocoder for audio conversion.



Paper Review 1:

Adversarial Neural Audio Synthesis

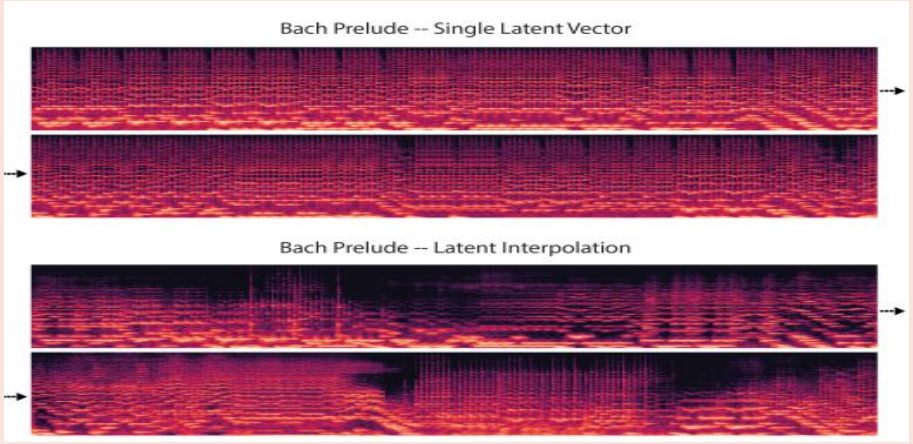
- Engel et al. (2019). *GANSynth: Adversarial Neural Audio Synthesis.
- GANs for high-quality, coherent instrument notes (NSynth).
- Key Ideas:
  - Model log magnitudes & Instantaneous Frequencies (IF) spectrally.
  - IF spectra > direct phase for coherence.
  - Higher STFT resolution = better performance (less blur).



Adversarial Neural Audio Synthesis

Adversarial Neural Audio Synthesis

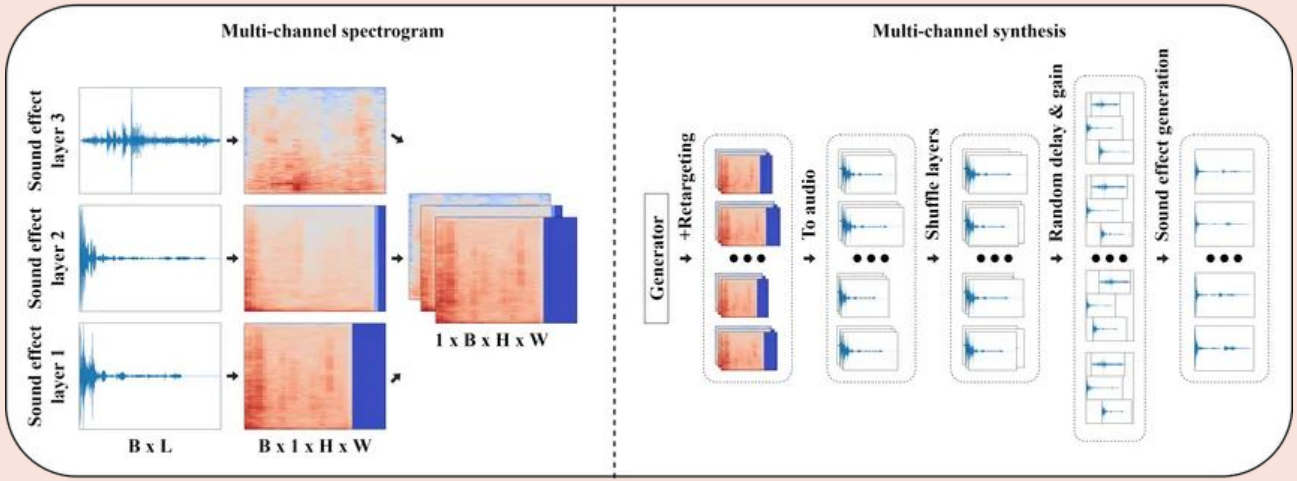
- outperformed a strong WaveNet baseline
- significantly faster generation
- Relevance to Thesis:
  - viable for high-quality audio synthesis
  - IF importance for GAN phase coherence.
  - Demonstrates pitch conditioning.



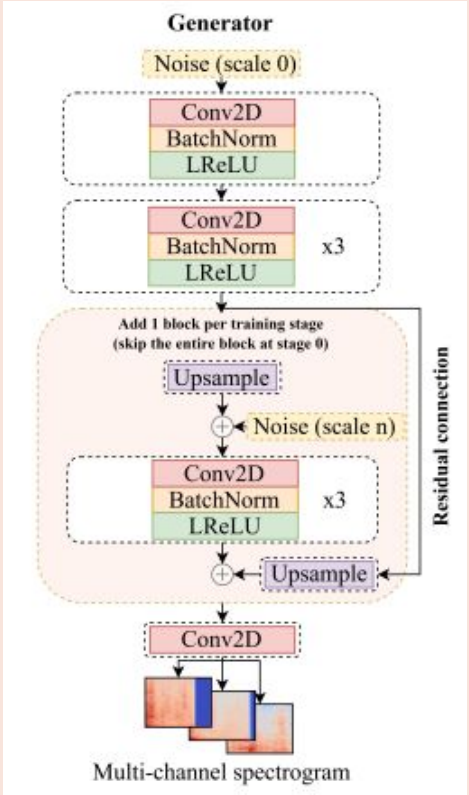
Paper Review 2:

Sound Effect Variation Synthesis

- Barahona-Ríos & Collins (2022). *SpecSinGAN: Sound Effect Variation Synthesis Using Single-Image GANs.*
- generate novel variations of a single one-shot sound (e.g., footstep, jump) from one example
- key Ideas & Contributions:
    - adapted single-image GANs (ConSinGAN) for audio
    - trained on multi-channel spectrograms
    - unconditional generation

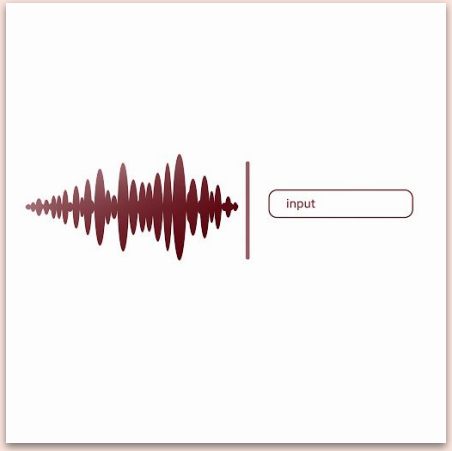


Sound Effect Variation Synthesis



Sound Effect Variation Synthesis

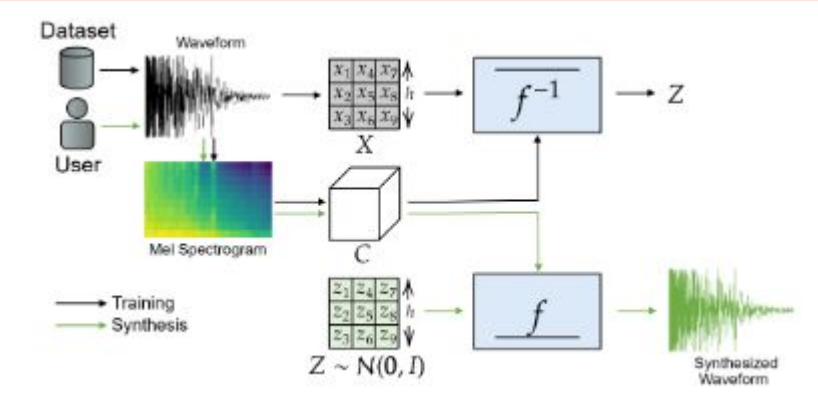
- produced plausible novel variations from a single sound
- outperformed baseline
- Relevance to Thesis:
  - addresses sound effect variation, reducing repetitiveness, enhancing realism
  - useful for scarce data scenarios
  - multi-channel spectrogram layers relevant for text-to-sound component specification



Paper Review 3:

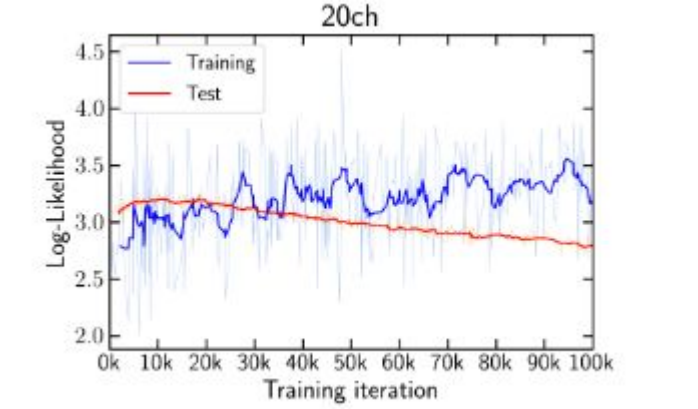
Neural Synthesis of Sound Effects with WaveFlow

- Andreu & Villanueva Aylagas (2022). *Neural Synthesis of Sound Effects Using Flow-Based Deep Generative Models.*
- Key Ideas & Contributions:
  - adopted WaveFlow (raw audio generative flow model) for sound effect synthesis
  - conditioned WaveFlow on a low-dimensional mel spectrogram of an example sound to guide generation and create variations
  - explored style transfer (e.g., explosions from percussive sound spectrograms).



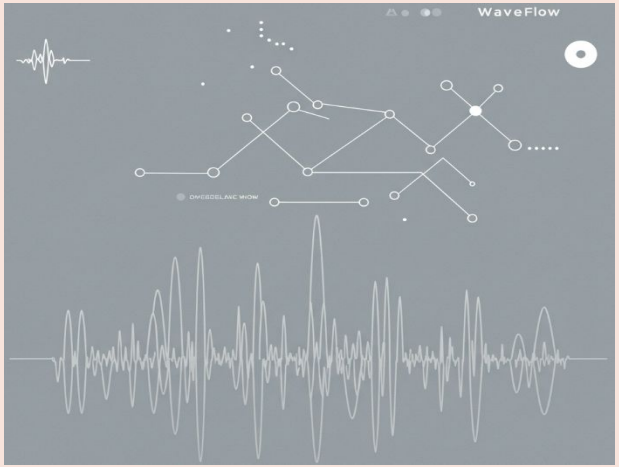
Neural Synthesis of Sound Effects with WaveFlow

- generation quality similar to training set with perceivable variations
- Identified trade-off: mel spectrogram affects quality vs. diversity
  - Lower dimension: more diversity, potentially lower fidelity
  - Higher dimension: better quality, less diversity



Neural Synthesis of Sound Effects with WaveFlow

- Relevance to Thesis:
- flow-based models effective for high-quality SFX generation
- shows controllable variation method using a conditioning signal
- style transfer hints at flexible sound generation, potentially guided by abstract text

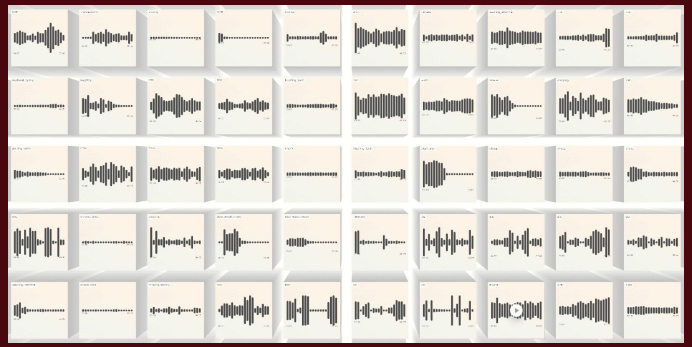


- GANSynth:
  - GANs + IF for high-quality, fast generation.
  - highlights spectral representation for coherence.
- SpecSinGAN:
  - single-example GAN + Multi-channel Spectrograms
  - addresses data scarcity, layered sound
- WaveFlow for SFX:
  - controllable SFX variations/style transfer
  - focus on direct waveform generation, conditioner impact

Summary

Problems with Current Solutions

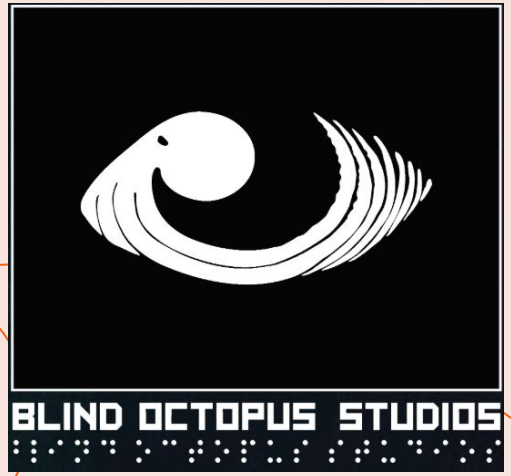
- **Data Bottleneck**: large (text, sound) datasets are scarce and hard to create.
- **General SFX Frontier**: direct text-to-SFX is an emerging research area.
- **Fine-Grained Control**: precisely controlling multiple sound attributes from text is difficult.
- **Trade-off Challenge**: balancing sound quality, diversity & text relevance is hard



Problems with Current Solutions

- **Generalization Issues**: models struggle with text descriptions unseen during training.
- **Evaluation Difficulties**: defining objective metrics for quality, diversity, and text relevance
- **VI User Needs Overlooked**: needs of visually impaired users

Bridging the Gaps

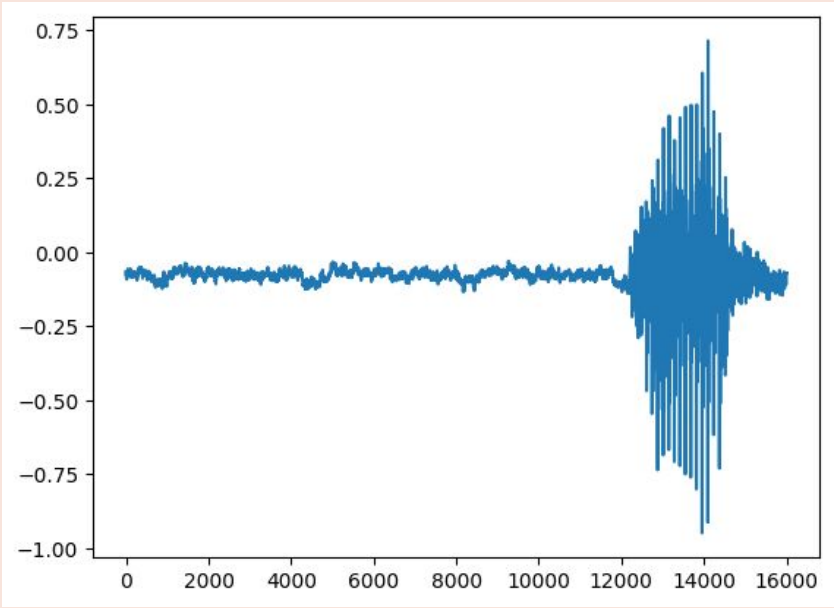


1. Pioneering Text-to-Sound for VI-Centric Game Audio
2. Embedding VI User Requirements
3. Exploring Controllable Variation & Nuance from Text
4. Developing Robust User-Centric Evaluation Methodologies

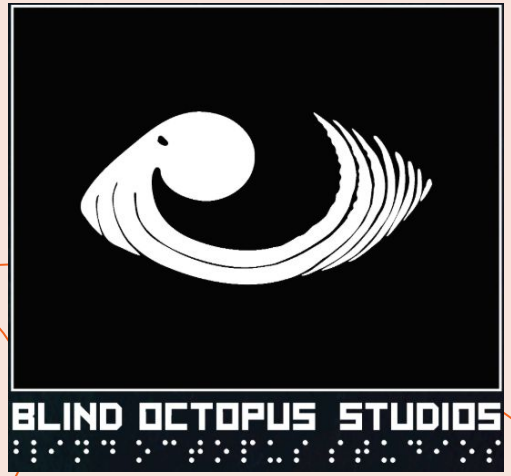
Implementation

```
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
import torchaudio
import sys

import matplotlib.pyplot as plt
import IPython.display as ipd

from tqdm import tqdm
```

```
plt.plot(waveform.t().numpy());
```



Future



1. Dataset Availability & Quality for Text-to-SFX
2. Effective Text-to-Sound Semantic Mapping
3. Meaningful & Scalable Evaluation for VI Users
4. Balancing Quality, Diversity, Controllability, Text-Relevance
5. Integrating User Feedback Systematically & Iteratively

Questions

Thank you!