U. PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Synthesizing Audio from Textual Input

Márcio Duarte
Luís Paulo Reis (Supervisor)
Gilberto Bernardes (Second Supervisor)

Faculdade de Engenharia da Universidade do Porto

Outline

- Introduction
    - Context
    - Background
    - Motivation
    - Research objectives
- State of the Art
    - Introduction
    - Traditional Soundscape Generation
    - Unsupervised Sound Generation
    - Vocoders
    - End-to-End Models
    - AudioLM
    - DiffSound
    - AudioGen
- Development
    - Datasets
    - Exploratory Experiments
    - GANmix

Overview of Computer Science and its Evolution

- “Computer Science is the study of computation and information.” [**university˙of˙york˙what˙nodate**]
- Evolution of Computer Science: From traditional programming to advanced Machine Learning (ML) and Deep Learning (DL) techniques.
- Importance of Big Data and Parallel Computing: Catalysts for advancements in ML and DL.
- Role of Deep Learning: Automated feature extraction, particularly effective in tasks like image and audio processing.
- Significance of Generative Models: Creating synthetic data for various applications.

Timeline of Key Events in AI and Machine Learning

1. **1646 - Birth of Leibniz:**
   - Gottfried Wilhelm Leibniz, the founder of computer science, is born.
2. **1958 - Birth of Modern AI:**
   - F. Rosenblatt proposes three fundamental questions leading to the development of the perceptron.
3. **1960s - Perceptron Convergence:**
   - Intensive work on convergence algorithms for the perceptron.
4. **1969 - First Winter of AI:**
   - Minksy and Papert demonstrate the limitations of perceptrons, leading to a slowdown in AI research.
   - This led to the first AI winter.
5. **1980s - Emergence of Multilayer Neural Networks:**
   - Studies on learning under multilayer neural networks.
6. **1986 - Backpropagation:**
   - Rumelhart et al. describe backpropagation, a key learning procedure for neural networks.
7. **1990s - Second Winter of AI:**
   - Decreased investments in ML due to lack of real successes.
8. **Turn of the Millennium - Resurgence of ML:**

Digital Audio Processing

- Sound can be discretized by sampling at a specific time rate, known as the sampling rate.
- Sampling rate impacts the accuracy of the signal, with the standard value being 44,100 Hz.
- Sound can be represented digitally as an array, enabling reconstruction by a computer.

Short-Time Fourier Transform (STFT):

- The STFT allows the analysis of the frequency content of a signal over time.
- It computes the Fourier Transform of the input signal within windowed time intervals.
- The resulting time-frequency representation is used to generate spectrograms.

Digital Audio Processing (contd.)

Spectrograms and Machine Learning:

- Spectrograms represent sound as a time-frequency image, enabling analysis and processing.
- Convolutional Neural Networks (CNNs) can be used for sound analysis, but filter shapes and convolution axes require careful consideration.

Soundscapes:

- Soundscapes are complex mixes of sounds heard in everyday life.
- They encompass natural, human-made, and cultural sounds, creating immersive auditory environments.
- Generating soundscapes with machine learning is challenging due to their lack of specific structure.

Foundations for Enhancing Generative Models for Audio

- **Data Augmentation:** Applying transformations to the original data to increase its size and diversity.
- **Evaluation Metrics:** Methods used to measure the quality and diversity of the generated sounds. They provide a way to compare different generative models and assess their strengths and weaknesses.
- **Data Embedding:** Converting data into numerical representations that capture its essential features and characteristics. This facilitates the learning process of generative models and enhances their expressiveness and efficiency.

Deep Learning Frameworks

- TensorFlow: offers a comprehensive ecosystem for building and deploying machine learning models, including support for audio-related tasks. TensorFlow provides a static computation graph and is known for its scalability and performance.
- PyTorch: has gained popularity due to its ease of use and dynamic computation graph, which allows for more flexibility and intuitive programming. PyTorch offers a user-friendly API and is known for its flexibility and support for various applications.
- Keras: Keras is a high-level deep learning library that runs on top of TensorFlow. It provides a user-friendly interface for building and training neural network models. Keras simplifies the creation of neural networks by abstracting away many low-level details, making it a popular choice for beginners.

In this thesis, we have selected PyTorch as the deep learning framework of choice due to its flexibility, ease of use, and optimal performance for our specific tasks.

Generative Deep Learning Architectures

- Generative deep learning architectures are designed to generate new and diverse data samples from a learned distribution.
- They learn to estimate the underlying probability distribution of the data by minimizing some distance or loss function between the model and the actual distribution.
- They have been applied to various tasks, such as image synthesis, text generation, and audio synthesis.
- In this thesis, we focus on some of the most popular and influential generative deep learning architectures, such as:
  - Deep autoregressive networks (DARNs)
  - Variational autoencoders (VAEs)
  - Generative adversarial networks (GANs)
  - Normalizing flows
  - Diffusion models
  - Transformers
  - Vector quantized variational autoencoders (VQ-VAEs)
  - Multi-scale vector quantized variational autoencoders (MS-VQ-VAEs)

Motivation

1. **Revolutionizing Sound Generation**
   - Integration of ML transforms sound creation and experience.
   - Opens creative avenues for artists, impacts industries like film, gaming, VR.
2. **Need for Further Research**
   - Urgency for studies in sound generation technologies.
   - This dissertation contributes significantly, offering resources for exploration.
3. **Impact and Contribution**
   - Reshaping human potential in sound creation through digital technologies.
   - Valuable resource for audio processing professionals, guiding future endeavors.

Research Objectives

1. Make a study of the current state-of-the-art deep learning architectures, focusing on generative ones.
2. Examine prior algorithms that can process sound for augmentation, feature extraction, or other purposes.
3. Make a study of the current state-of-the-art architectures used to develop sounds artificially.
4. Develop end-to-end systems that can synthesize sound from any given text input, while accounting for hardware constraints and ensuring reliable performance.
5. Evaluate the systems' ability to generate a sound from the given textual input accurately.

State of the Art

- There are different types of sound generation methods, depending on the level of abstraction and supervision involved
- Four types of sound generation methods are considered in this thesis:
  - Traditional
  - Unsupervised
  - Vocoders
  - End-to-end models

Traditional Soundscape Generation

Notable Tools

- *Scaper* [**salamon˙scaper˙2017**]: Open-source library for synthetic sound environments
- SEED [**bernardes˙seed˙2016**]: System for resynthesizing environmental sounds with precise control over variation
- Physics-Based Concatenative Sound Synthesis [**magalhaes˙physics-based˙2020**]: Creates novel auditory experiences by assembling pre-existing sound segments

Unsupervised Sound Generation

Approach

- Learn sound features and distributions without explicit labels
- Utilize unlabeled audio data for pattern capture and structure learning
- Valuable when labeled datasets are limited or costly

Notable Models

- WaveGAN [**donahue˙adversarial˙2019**]: Unsupervised waveform synthesis using modified GAN
- Generative Transformer [**verma˙generative˙2021**]: Autoregressive prediction of audio samples using transformer networks
- wav2vec 2.0 [**baevski˙wav2vec˙2020**]: Speech generation model with convolutional feature encoder, Transformer, and quantization module
- SoundStream [**zeghidour˙soundstream˙2021**]: Neural audio codec for efficient audio compression

Vocoders

Notable Models

- WaveNet [**oord˙wavenet˙2016**]: Generative neural network using dilated causal convolutions for raw audio waveform generation
- WaveNet Variants: Models like WaveRNN, FloWaveNet, and Fast WaveNet reduce complexity while maintaining effectiveness
- MelGAN [**kumar˙melgan˙2019**]: GAN-based model using Mel-Spectrograms for coherent audio waveform generation
- GANSynth [**engel˙gansynth˙2019**]: GAN using log-magnitude spectrograms and phases for waveform generation
- HiFi-GAN [**kong˙hifi-gan˙2020**]: GAN model combining efficiency and high-fidelity speech synthesis

End-to-End Models

Introduction

- Traditional audio synthesis involves multiple stages
- Challenges: expertise, design choices, errors, inconsistencies
- End-to-end models directly map text to audio waveform with neural networks

Frameworks

- Specialized models for specific domains (speech, music)
- Universal models for broader applications

End-to-End Audio Models Comparison

Table: A comparison of different end-to-end generative models for audio.

Model	**Type**	**Input**	**Output**
Char2wav [**sotelo˙char2wav˙2017**]	Speech	Text prompt	Raw audio waveform
VALL-E [**wang˙neural˙2023**]	Speech	Text and acoustic prompt	Raw audio waveform
Jukebox [**dhariwal˙jukebox˙2020**]	Music	Genre, artist, and lyrics	Raw audio waveform
Riffusion [**forsgren˙riffusion˙2022**]	Music	Text prompt	Raw audio waveform
MusicLM [**agostinelli˙musiclm˙2023**]	Music	Text prompt	Raw audio waveform
SampleRNN [**mehri˙samplernn˙2017**]	General	None	Raw audio waveform
AudioLM [**borsos˙audiolm˙2022**]	General	Text prompt	Raw audio waveform

Text-to-Speech (TTS)

Definition

- Convert written text into synthesized speech
- Use deep neural networks for direct mapping
- Notable TTS Models:
  - Char2wav
  - VALL-E

Generative Music

Definition

- Create music using generative techniques
- End-to-end models for composing new musical pieces
- Notable Generative Music Models:
  - Jukebox
  - Riffusion
  - MusicLM

General Text-to-Audio

Definition

- Convert various forms of text to corresponding audio outputs
- Applications: sound effects, voice transformation, environmental sound synthesis
- Notable Text-to-Audio Models:
  - SampleRNN
  - AudioLM
  - DiffSound
  - AudioGen

AudioLM

- A framework for high-quality audio generation with long-term consistency [**borsos˙audiolm˙2022**].
- Maps input audio to a sequence of discrete tokens and treats audio generation as a language modeling task.
- Achieves high-quality synthesis and long-term structure through a hybrid tokenization scheme of semantic and acoustic tokens.
- Consists of three main components: tokenizer, language model, and detokenizer.
- Generates syntactically and semantically plausible speech and music continuations without any transcript or annotation.

DiffSound

AudioGen

Dataset Introduction

- Datasets form the foundation for generative models
- Learning material for DL algorithms and evaluation of audio outputs
- Two types of datasets: categorical and descriptive

Table of Datasets

Table: Comparison of datasets for soundscapes

Name	Type	# Samples	Duration	Labels
Acoustic Event Dataset [**takahashi˙deep˙2016**]	Categorical labeled	5223	Average 8.8s	One of 28 labels
AudioCaps [**kim˙audiocaps˙2019**]	Descriptive labeled	39597	10s each	9 words per caption
AudioSet [**gemmeke˙audio˙2017**]	Categorical labeled	2084320	Average 10s	One or more of 527 labels
Audio MNIST [**becker˙interpreting˙2018**]	Categorical labeled	30000	Average 0.6s	One of 10 labels
Clotho	Descriptive	4981	15 to 30s	24 905

Exploratory Experiments

- Objective: Gain insights for GANmix development
- Experiments: Classification, GAN, AE, VAE
- Findings: Foundation for audio representation, GAN effectiveness, AE/VAE capabilities
- Impact: Crucial for robust audio generation with GANmix

GANmix: Introduction

- GANmix: Fusion of GAN and VAE for audio generation under constraints.
- Addresses computational limitations for high-quality audio.
- Combines GAN's generative power with VAE's latent space manipulation.

GANmix: Model Architecture

- Generator and discriminator operate in latent space.
- VAE Training: Computational challenge, requires extensive datasets.
- AudioLDM's High-Performance: Top model for audio generation.
- Accessibility of AudioLDM: Open source, accessible via Hugging Face's model hub.

GANmix: Experimental Results

- Preliminary experiments with Audio MNIST: Promising but suboptimal.
- Refinements: Different optimizers, model sizes, loss functions.
- Clotho dataset: Significant improvement in generated audio quality.
- Challenges in achieving equilibrium between generator and discriminator.

GANmix: Final Model

- GANmix architecture with Clotho dataset: Significant improvements.
- Unlike typical models using CNN, GANmix uses fully connected neural networks.
- Generator input: Random Gaussian noise, passes through hidden layers.
- Discriminator: Takes embedding as input, applies tanh activation.
- Loss function: BCE. Optimized with Adam. Learning rate updates every 10 epochs.

Results

- Setup: Overview of experimental conditions and configurations.
- Presentation of Results: Showcase of outcomes from GANmix experiments.
- Discussion: Analyzing and interpreting the obtained results.
- Constraints and Challenges: Addressing limitations and difficulties encountered.

Experimental Setup

- GANmix model trained using BCE loss in PyTorch.
- Utilized two datasets: Audio MNIST and Clotho for diverse training.
- Three hardware setups: Kaggle, LIACC 1, LIACC 2, tailored for resources.
- Implementation: Python, PyTorch framework.
- Preprocessing: Randomly cropped samples to 5 seconds for diversity.
- Hyperparameters adjusted for batch size, epochs, and learning rates.
- Stopped training based on convergence for resource efficiency.

Presentation of Results

- Objectives: Explore generative AI models for audio production, assess performance.
- Context: Each experiment designed with specific research questions and hypotheses.
- Methodology: Hardware, software, and architecture details provided for transparency.
- Evaluation: Performance assessed through evolving loss plots and spectrograms.
- Systematic Organization: Ensures a comprehensive understanding of procedures and results.
- Basis for Analysis: Provides foundation for discussing effectiveness of generative AI models.

Experiment X: Title

- Objectives: [Objectives of the experiment]
- Model Details: [Describe key details of the model used, e.g., parameters, loss function]
- Dataset: [Mention the dataset used for training and evaluation]
- Optimizer and Learning Rate: [Specify the optimizer and learning rate used]
- Training Process: [Provide essential details about the training process, e.g., convergence status]

[Optional: Any unique aspects or considerations for this experiment]
[Optional: Any additional insights or observations from this experiment]

Analysis and Interpretation

- Identifying Trends
- Results for Future Investigation
- Interpretation of Results
- Conclusion

Identifying Trends

- Inverse correlation between generator and discriminator losses
- Convergence tends to plateau after a certain number of epochs
- Impact of learning rate on convergence speed
- Influence of optimization algorithms (e.g., SGD, RMSprop, Adam)
- Benefits of regularization methods (e.g., dropout, batch normalization, Gaussian noise)
- Importance of dataset size
- Exploration of latent space

Results for Future Investigation

- Occurrence of performance decline and NaN losses in certain experiments
- Further exploration of elastic network regularization
- Investigation of continuously increasing generator loss

Interpretation of Results

- Results didn't meet initial expectations but show potential
- Latent space exploration as a promising strategy
- Limitations of small datasets, especially in audio length and quantity
- Need for access to comprehensive datasets
- Computational resource challenges

Conclusion

- Analysis and interpretation of trends and patterns
- Potential for future advances in generative AI models for audio synthesis
- Lack of satisfactory practical results due to dataset limitations
- Importance of comprehensive datasets and computational resources

Constraints and Challenges

- Hardware Resources
- Data Quality and Quantity
- Hyperparameter Tuning

Hardware Resources

- Scarcity of hardware resources for training and evaluation
- Challenges in accessing sufficient computing power and memory
- Strategies adopted to optimize hardware usage
- Impacts, trade-offs, and opportunities resulting from resource limitations

Data Quality and Quantity

- Challenges posed by the quality and quantity of available data
- Importance of high-quality and diverse data for generative models
- Strategies employed to mitigate data limitations
- Considerations regarding data augmentation techniques

Hyperparameter Tuning

- Time constraints and challenges in hyperparameter tuning
- Significance of hyperparameters in model performance
- Impact of default or arbitrary values on model potential
- Recommendations for future work in hyperparameter optimization

Conclusion

- Discussion of major limitations and challenges faced in solution development
- Description of strategies employed to address these issues
- Possible implications, trade-offs, and opportunities arising from constraints
- Affirmation of the proposed solution's strengths and advancements in generative AI models for audio synthesis

Overview and Reflections

- Comprehensive Study of State-of-the-Art Deep Learning Architectures for Audio Synthesis
- Development of End-to-End Systems for Sound Synthesis and Evaluation
- Challenges and Lessons Learned

Future Directions

- Exploring Novel Architectures
- Dataset Expansion
- Evaluation Metrics

Novel Architectures

- Briefly introduce the proposed theoretical architectures
- Mention their objectives, design principles, and potential applications
- Highlight the need for future research and development in this area

Conclusion

- Summarize the main achievements and contributions of the research
- Emphasize the progress made in understanding and developing generative AI models for audio synthesis
- Acknowledge the challenges and ongoing work required for further advancements