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

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. **1958 - Birth of Modern AI:**
   - F. Rosenblatt proposes three fundamental questions leading to the development of the perceptron.
2. **1960s - Perceptron Convergence:**
   - Intensive work on convergence algorithms for the perceptron.
3. **1969 - Limitations of Perceptrons:**
   - Minksy and Papert demonstrate the limitations of perceptrons, leading to a slowdown in AI research.
4. **1980s - Emergence of Multilayer Neural Networks:**
   - Studies on learning under multilayer neural networks.
5. **1986 - Backpropagation:**
   - Rumelhart et al. describe backpropagation, a key learning procedure for neural networks.
6. **1990s - Second Winter of AI:**
   - Decreased investments in ML due to lack of real successes.
7. **Turn of the Millennium - Resurgence of ML:**
   - Emergence of three trends: Big Data, reduced cost of parallel computing, and interest in Deep Neural Networks (DNN).
8. **2010s - DL in Everyday Applications:**
   - DL becomes integral for various computer-made tasks,

Introduction to Background

Digital Audio Processing 1

- 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 2

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

To develop generative models for audio, it is necessary to address several factors that impact their performance and quality. This thesis concentrates on three main areas:

- **Data Augmentation:** Applying transformations to the original data to increase its size and diversity. This helps overcome limitations of small or imbalanced datasets and improves the generalization ability of generative models.
- **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

Deep learning frameworks have played a pivotal role in advancing the field of AI, enabling researchers and practitioners to efficiently develop and deploy complex neural networks. These frameworks provide a wide range of tools and techniques for building, training, and evaluating deep neural networks, accelerating progress in the field.

- TensorFlow: TensorFlow is a widely-used deep learning framework developed by Google. It 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: PyTorch is an open-source deep learning framework developed by Facebook. It 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.

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. **ML in Audio Processing**
   - ML techniques enhance sound synthesis, restoration, and speech recognition.
   - Learn complex patterns, improving quality and efficiency.
2. **Revolutionizing Sound Generation**
   - Integration of ML transforms sound creation and experience.
   - Opens creative avenues for artists, impacts industries like film, gaming, VR.
3. **Need for Further Research**
   - Urgency for studies in sound generation technologies.
   - This dissertation contributes significantly, offering resources for exploration.
4. **Impact and Contribution**
   - Reshaping human potential in sound creation through digital technologies.
   - Valuable resource for audio processing professionals, guiding future endeavors.
5. **Overall Significance**
   - Valuable contribution to audio processing and machine learning field.

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.

Introduction

- Sound generation is the task of creating realistic and expressive sounds from scratch or based on some input
- It has many applications in music, entertainment, education, and research
- 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

Approach

- Threefold strategy: Segmentation, Feature extraction, Resynthesis
- Statistical models and stochastic processes used

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