Audio Generation with Foundation Models – Handout

Einführung

Audio-Foundation-Modelle funktionieren ähnlich wie große Sprach- oder Bildmodelle: Sie lernen Muster in ihren Input-Daten, komprimieren diese in handhabbare Repräsentationen und wandeln ihre Vorhersagen wieder in hörbares Audio um. Modelle dieser Art können Musik, Sprache oder Soundeffekte erzeugen, ohne dass sie explizit mit Labels trainiert werden müssen.

Im Vergleich zu Text- oder Bildmodellen ist Audio besonders anspruchsvoll, da es kontinuierlich abläuft und in vielen Dimensionen strukturiert ist. Mehrere Faktoren beeinflussen die Audio-Generierung:

- **Zeitliche Abhängigkeit:** Audio verändert sich kontinuierlich über die Zeit. Ein Modell muss die Konsistenz über die gesamte Sequenz bewahren, sonst entstehen sofort hörbare Brüche oder Unstimmigkeiten.
- **Hierarchische Struktur:** Audio besitzt viele Ebenen – von rhythmischen Mustern über Harmonien bis zu feinen Klangdetails. Diese Ebenen müssen miteinander abgestimmt sein, damit Musik, Sprache oder Effekte sinnvoll klingen.
- **Hohe Datenkomplexität:** Audiosignale enthalten sehr viele Informationen pro Zeiteinheit. Das macht die Modellierung aufwendig und erfordert spezielle Ansätze, um diese Komplexität zu beherrschen.
- **Perzeptive Sensibilität:** Kleine numerische Fehler oder Ausreißer können sofort als Knackser, Verzerrungen oder unnatürliche Artefakte hörbar werden.

Deswegen unterscheiden sich Audio-Modelle in Architektur und Trainingsansatz von Bild- oder Sprachmodellen und müssen dementsprechend angepasst werden.

Technologien & Prinzipien

Wichtige technologische Bausteine welche die hier vorgestellten Modelle ausmachen:

- **1D-Convolutions:** Diese filtern Audiosignale entlang der Zeitachse, erkennen wiederkehrende Muster, rhythmische Strukturen oder spektrale Veränderungen. Sie bilden die Basis vieler GAN- und Transformer-Modelle, die direkt auf Roh-Audio arbeiten.
- **VQ-VAE / SoundStream**: Diese Autoencoder komprimieren Audiodaten in diskrete Tokens. So werden lange Sequenzen handhabbar und die wesentlichen Strukturen des Signals können von einem Modell gelernt, und dessen Details später rekonstruiert werden. Zusätzlich kann VQ-VAE grob die semantische Struktur von der akustischen Detailinformation trennen.
- **CLAP-Embeddings:** Sie schaffen einen gemeinsamen Repräsentationsraum für Text und Audio. Dadurch kann Audio gezielt über Text-Prompts gesteuert werden, ohne dass Text-Audio Verhältnisse im Trainingsset gepaart vorhanden sein müssen.

Diese Technologien bilden die Grundlage für Transformer-, Diffusions- und GAN-basierte Modelle, die auf großen Audio-Datensätzen trainiert werden und die Basis moderner Musik-, Sprach- und Effekt-Generierung darstellen.

Modelle & Funktionsweise

Die vorgestellten Audio-Foundation-Modelle zeigen die Entwicklung von einfachen Soundgeneratoren hin zu universellen Systemen auf, die komplexe, langzeitkohärente Audiosequenzen erzeugen können.

WaveGAN (2018)

- Input: Kurze Audio-Clips (Drum-Hits, Sprache, Soundeffekte)
- Output: Neue kurze Waveform-Sounds
- Aufbau: Generative Adversarial Network
    - Generator: transponierte 1D-Convolutions erzeugen neue Wellenformen
    - Diskriminator: 1D-Convolutions unterscheiden Real vs. Fake

Das Modell lernt direkt auf Roh-Audioebene und erzeugt realistisch klingende Samples. Der Fokus liegt auf kurzen Clips, die sich gut für Sounddesign oder Sprach- und Effektclips eignen.

Jukebox (2020)

- Input: Roh-Audio kompletter Songs
- Output: Ganze Songs mit Gesang, Instrumenten und möglicher Stil-/Genreanpassung
- Aufbau: Autoregressiver Transformer mit VQ-VAE-Kompression
    - Roh-Audio wird in diskrete Tokens umgewandelt
    - Transformer lernt musikalische Strukturen und Zusammenhänge zwischen Tokens

Das Modell generiert Musik tokenweise, wodurch längere, hierarchisch strukturierte Songs entstehen können.

AudioLM (2022)

- Input: Große Audio-Datasets (Musik & Sprache)
- Output: Langfristig konsistente Audiosequenzen
- Aufbau: Autoregressiver Transformer auf zwei Token-Ebenen
    - Semantische Tokens: beschreiben Bedeutung und Struktur
    - Akustische Tokens: beschreiben Klangdetails

Das Modell lernt sowohl zeitliche als auch hierarchische Zusammenhänge und bewahrt die Konsistenz über lange Sequenzen hinweg, wodurch realistische Musik- und Sprachverläufe entstehen.

AudioLDM (2023)

- Input: Text-Audio Embeddings (CLAP)
- Output: Text-gesteuerte Audioerzeugung (Musik, Sprache, Effekte)
- Aufbau: Latent-Diffusionsmodell
    - Diffusionsprozess: startet mit Rauschen und wird schrittweise zu realistischem Audio „denoised“
    - CLAP-Embeddings ermöglichen flexible Textsteuerung

Durch den Diffusionsprozess können Musik, Sprache oder Soundeffekte direkt aus Text-Prompts generiert werden. Das Modell ist besonders flexibel und kann unterschiedliche Audioarten erzeugen.

Entwicklung der Modelle

Die Modelle bauen aufeinander auf, übernehmen bewährte Konzepte und versuchen, frühere Einschränkungen zu verbessern:

- **WaveGAN:** Ansatz für Roh-Waveforms und GAN-basierte kurze Sequenzen.
- **Jukebox:** Übernimmt generative Idee und erweitert sie mit Tokenisierung und Transformer-Architektur.
- **AudioLM:** Verfeinert Token-Ansatz, trennt semantische und akustische Tokens für bessere Langzeitkonsistenz.
- **AudioLDM:** Übernimmt komprimierte Audio-Representationen, ersetzt den Transformer durch ein Diffusionsmodell für flexible Textsteuerung.

Zukunft der Audiogenerierung

Zukünftige Modelle können sich auf folgende Aspekte konzentrieren und diese weiter ausbauen:

- Höhere wahrgenommene Audioqualität durch präzisere Tokenisierung und verbesserte Signalverarbeitung.
- Echtzeitgenerierung und interaktive Steuerung über Text oder andere Eingabemodalitäten.
- Weiterentwicklung hin zu Systemen, die jede Art von Klang verstehen und erzeugen können, einschließlich komplexer musikalischer oder sprachlicher Strukturen.

Bei der Weiterentwicklung von Audiogenerierung müssen einige Hürden beachtet werden:

- “Ghost Artifacts”, kleine, störende Klangfehler, die durch Ungenauigkeiten im Modell entstehen und leicht wahrnehmbar sind.
- Hoher Rechenaufwand und große Datenmengen, sowohl fürs Training als auch für die Echtzeitgenerierung.
- Fehlende Feinkontrolle und Schwierigkeit, Stil, Emotion oder räumliche Tiefe gezielt zu Steuern und Darzustellen.

Quellen: **WaveGAN**: ADVERSARIAL AUDIO SYNTHESIS (2018 / UC San Diego [Yi-Jun Tsai])**; OpenAI Jukebox**: A Generative Model for Music (2020 / OpenAI [Chris Donahue]); **AudioLM**: a Language Modeling Approach to Audio Generation (2022 / Google Research / DeepMind [Prafulla Dhariwal]); **AudioLDM**: Text-to-Audio Generation with Latent Diffusion Models (2023 / University of Surrey (UK) [Zalán Borsos]); A survey of deep learning audio generation methods (2024 / National Yang Ming Chiao Tung University (Taiwan) [Haohe Liu]); **ChatGPT** (GPT-5.1, OpenAI, 2025)