AUDIO GENERATION WITH FOUNDATION MODELS

The Velvet Sundown released two albums before admitting their music, images and backstory were created by AI

AI-generated image of AI-generated band the Velvet Sundown playing AI-generated music. Illustration: Velvet Sundown

https://www.theguardian.com/technology/2025/jul/14/an-ai-generated-band-got-1m-plays-on-spotify-now-music-insiders-say-listeners-should-be-warned

https://open.spotify.com/artist/2GRtyAXWUiisGYub5SGMrb

GLIEDERUNG

- Audio-Generierung
- Modelle & Architekturen:
  - Technologien & Prinzipien
  - WaveGAN
  - OpenAI Jukebox
  - AudioLM
  - AudioLDM
- Entwicklung
- Anwendungen
- Zukunft

AUDIO-GENERIERUNG

- **Zeitliche Abhängigkeit:** Audio ändert sich kontinuierlich - muss im Inhalt konsequent bleiben.
- **Hierarchische Struktur:** Komposition, Rhythmus, Klangdetails usw. müssen aneinander angepasst sein.
- **Perzeptive Sensibilität:** Numerische Fehler & Ausreißer sind als Knackser, Verzerrungen etc. leicht bemerkbar.
- **Hohe Datenkomplexität:** Komplexe Struktur, die andere Modellierungsansätze erfordert.

MODELLE

WaveGAN	Jukebox	AudioLM	AudioLDM
▪ **Generative Adversarial Network Model**	▪ **Autoregressive-Transformer-Model**	▪ **Language-Model Prinzip** auf Audio angewandt	▪ **Latent-Diffusion-Model**

TECHNOLOGIEN & PRINZIPIEN

1D Convolutions

- Analysieren Audiosignale entlang der Zeitachse, „scannen“ das Signal und finden wiederkehrende Muster/Strukturen

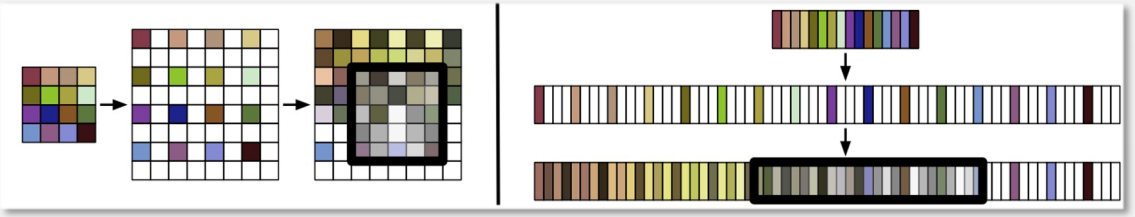
VQ-VAE/Soundstream

- Komprimieren von Audiodaten zu Tokens zur Verarbeitung & Informations-Abstrahierung

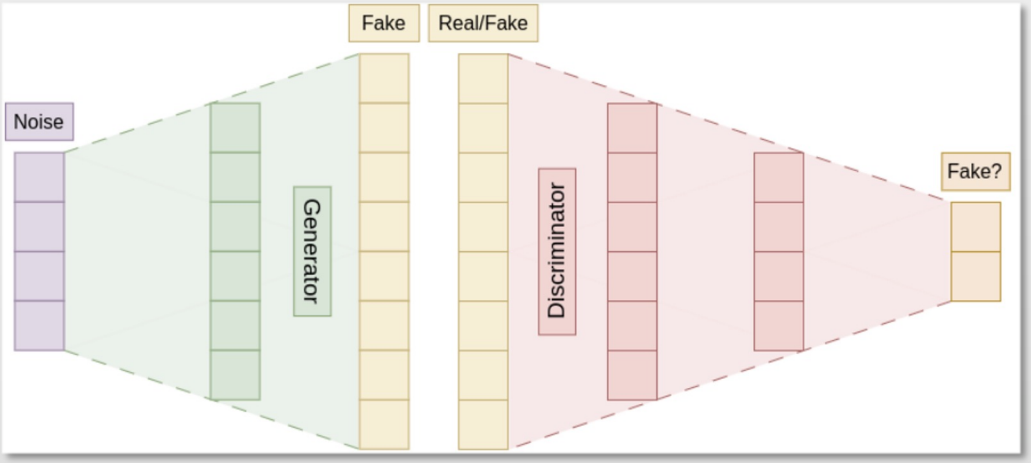
CLAP-Embeddings

- Gemeinsamer Repräsentationsraum für Text und Audio für textgesteuerte Audio-Generierung

WAVEGAN - 2018



- **Generative Adversarial Network Model,** welches direkt mit rohen Audiodaten in Waveform arbeitet
- **Generator:** transponierte 1D Convolutions
- **Discriminator:** 1D Convolutions

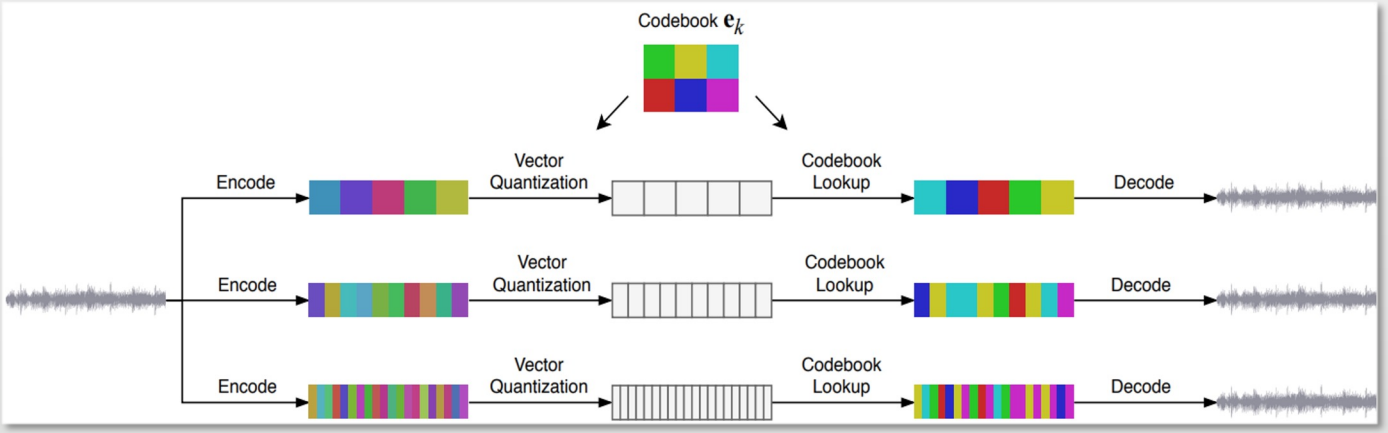


Generiert neue, realistische Audio Samples, geeignet für Sounddesign, Sprach- & Effectclips

OPENAI JUKEBOX - 2020

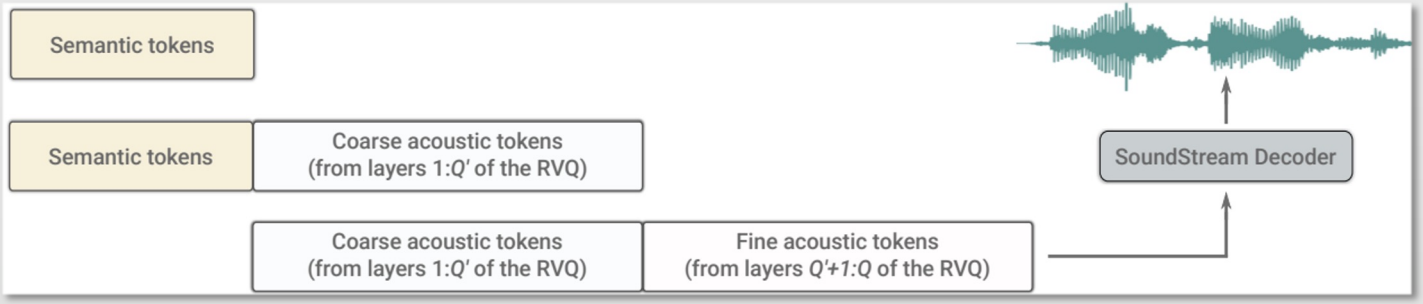
- **Autoregressives Transformer-basiertes Modell,** arbeitet mit rohen Audiodaten und Transformern
- **VQ-VAE-Kompression:** Wandelt Roh-Audio in diskrete Tokens um
- **Transformer:** „Lernt“ musikalische Strukturen innerhalb der Token-Sequenzen
- **Generierung:** Token für Token Umwandlung in Audiosignal

Generiert komplette Songs mit Gesang, Instrumenten und möglicher Genre / Stil-Anpassung



AUDIOLM - 2022

- **Language-Model Prinzip** auf Audio Generierung angewandt
- **Semantische Tokens** – Bedeutung & Struktur
- **Akustische Tokens** – Klangdetails
- **Transformer: „**Lernt“ zeitliche und hierarchische Zusammenhänge der Token-Sequenzen
- **Generierung:** Autoregressiv, basierend auf semantischen & akustischen Tokens

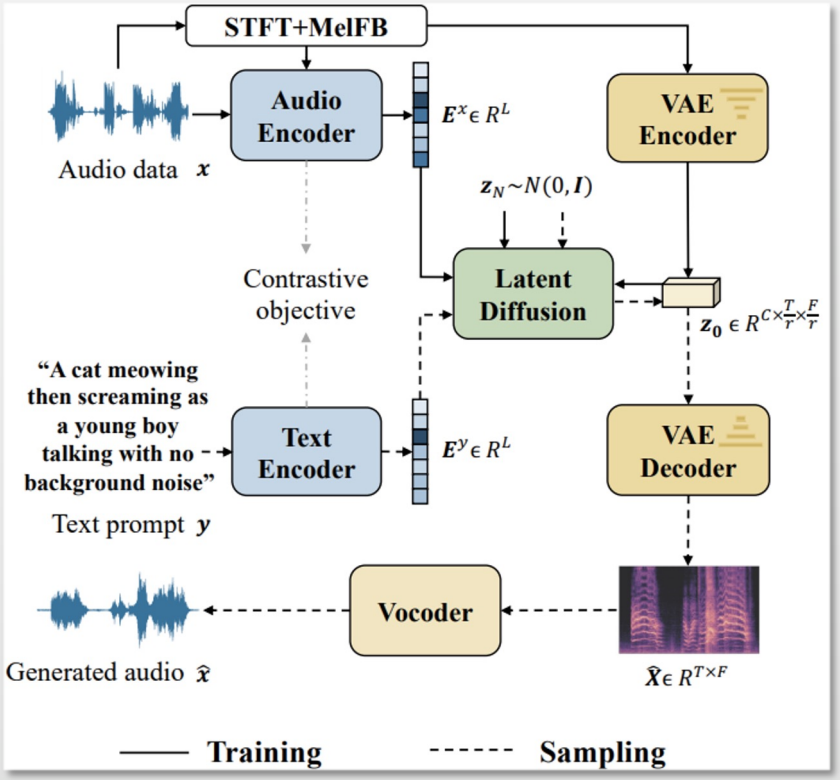


Generiert realistische Audio-Sequenzen, bewahrt Struktur und Konsistenz auch auf Langzeit

AUDIOLDM - 2023

- **Latent-Diffusion-Model Prinzip** auf Audio-Generierung angewandt
- **CLAP Embeddings:** Verbinden Text & Audio und ermöglichen flexible Prompt-Steuerung
- **Diffusionsprozess:** Beginnt mit Rauschen und wird schrittweise zu realistischem Audio “denoised”

Generiert Text-gesteuerte Musik, Sprache & Soundeffekte anhand von Textprompts



ENTWICKLUNG

WaveGAN
- Ansatz für Audiogenerierung mit rohen Waveforms & zeitlichen Sequenzen

Jukebox
- Übernimmt generative Audioidee / Erweitert mit Tokens & Transformer

AudioLM
- Übernimmt Token-Ansatz/ Erweitert mit Aufteilung in semantisch & akustisch

AudioLDM
- Übernimmt komprimiertes Audio / Ersetzt Transformer-Modell mit Diffusionsmodell

VIDEO TO AUDIO

https://replicate.com/zsxkib/mmaudio/examples#5j3c8p820hrmc0ckpvf8sctd84

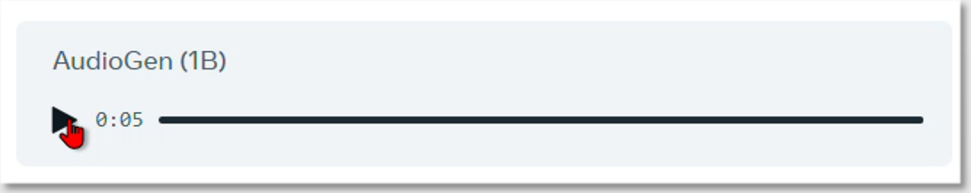
TEXT TO AUDIO / SPEECH TO AUDIO

https://elevenlabs.io

Input:

"Sirens and a humming engine approach and pass"

Output:



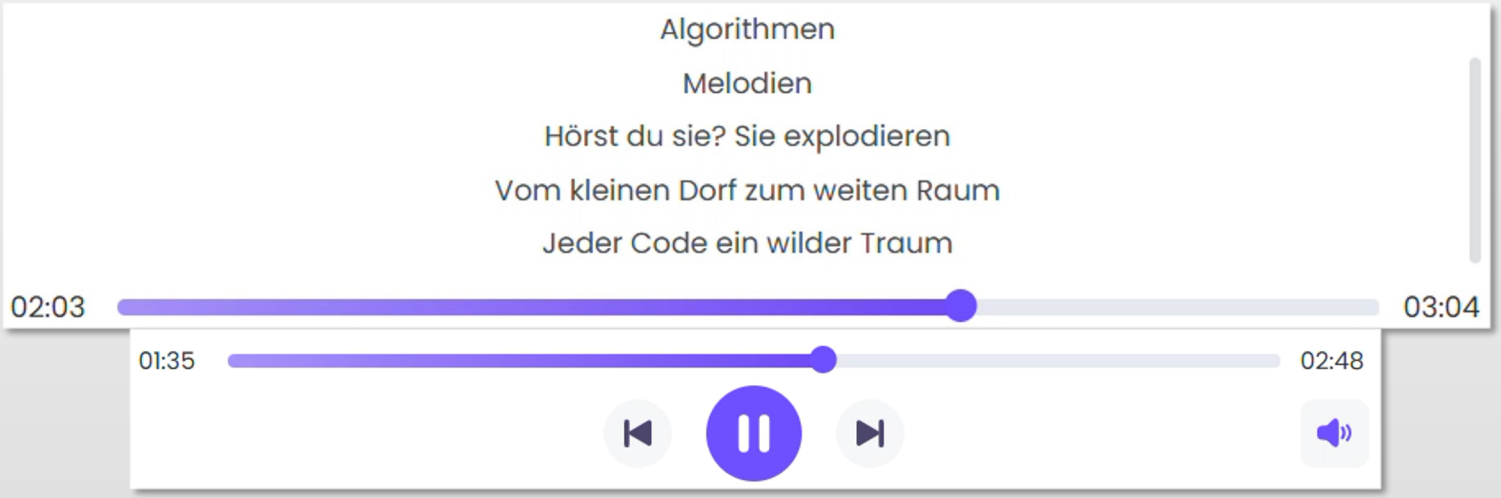
https://audiocraft.metademolab.com/audiogen.html

MUSIK-GENERIERUNG

Input:

"Ein Lied über Audio generierende Foundation models und die HFU in Furtwangen, fröhlicher, energetischer Rock / Glam Track“

Output:



https://de.topmediai.com/app/ai-music/?language=de

ZUKUNFT

Ziele & Entwicklungen

- Bessere wahrgenommene Qualität
- Weiterentwicklung für komplexeren Input & Output
- Echtzeitgenerierung & interaktive Steuerung

Hürden

- Hoher Rechenaufwand für Training & Generierung
- „Ghost Artifacts“: Klangfehler durch Ungenauigkeiten
- Fehlendes Finetuning in der Steuerung

QUELLEN

- **WaveGAN**: ADVERSARIAL AUDIO SYNTHESIS (2018 / UC San Diego [Yi-Jun Tsai])

https://arxiv.org/pdf/1802.04208

- **OpenAI Jukebox**: A Generative Model for Music (2020 / OpenAI [Chris Donahue])

https://arxiv.org/pdf/2005.00341

- **AudioLM**: a Language Modeling Approach to Audio Generation (2022 / Google Research / DeepMind [Prafulla Dhariwal])

https://arxiv.org/pdf/2209.03143

- **AudioLDM**: Text-to-Audio Generation with Latent Diffusion Models (2023 / University of Surrey (UK) [Zalán Borsos])

https://arxiv.org/pdf/2301.12503

- A survey of deep learning audio generation methods (2024 / National Yang Ming Chiao Tung University (Taiwan) [Haohe Liu])

https://arxiv.org/pdf/2406.00146

- **ChatGPT** (GPT-5.1, OpenAI, 2025)

https://chatgpt.com

BILD/VIDEO QUELLEN

- https://www.theguardian.com/technology/2025/jul/14/an-ai-generated-band-got-1m-plays-on-spotify-now-music-insiders-say-listeners-should-be-warned
- https://open.spotify.com/artist/2GRtyAXWUiisGYub5SGMrb
- https://replicate.com/zsxkib/mmaudio/examples#5j3c8p820hrmc0ckpvf8sctd84
- https://elevenlabs.io
- https://audiocraft.metademolab.com/audiogen.html
- https://de.topmediai.com/app/ai-music/?language=de