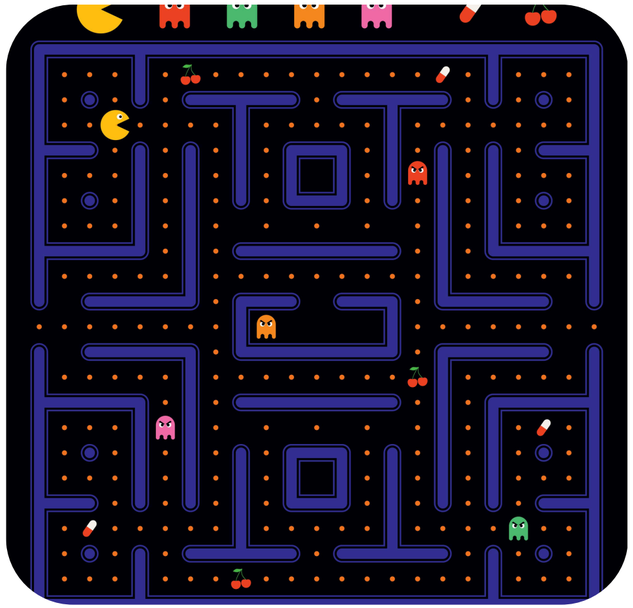
AUDIO SIGNAL ANALYSIS WITH WORLD MODELS

IA 376 2025.2
Davi Pincinato 157810
Henrique Parede de Souza 260497
Isadora Minuzzi Vieira 290184
Raphael Carvalho da Silva e Silva 205125

HOW DOES AN AI LEARN TO WIN A GAME

?



BY PLAYING MILLIONS OF TIMES, AN AI LEARNS THE PATTERNS AND PHYSICS OF THE GAME WORLD

1.

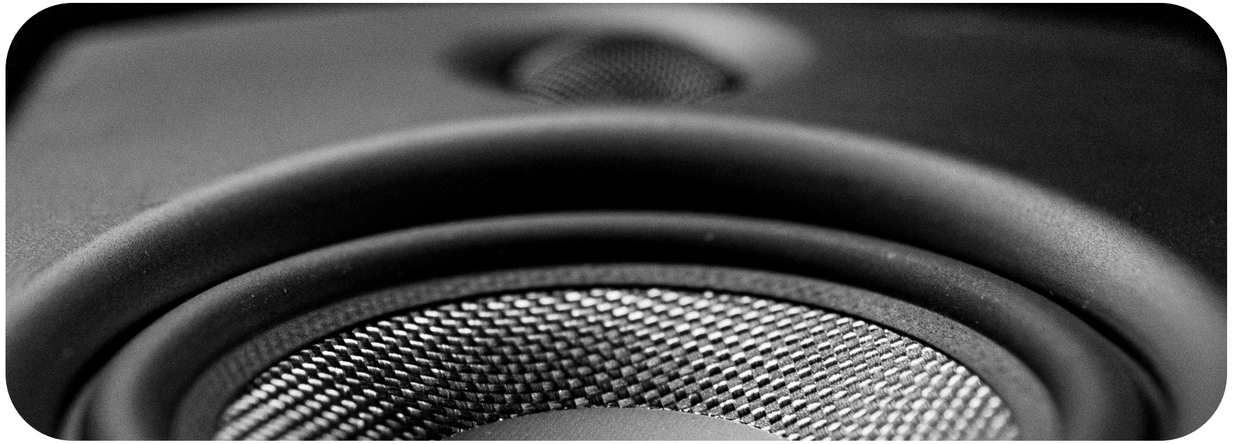
BUILDING AN INTERNAL WORLD

The AI’s simulation

An AI’s internal simulation of an environment’s rules, allowing it to predict future events

WORLD MODEL

THIS IS THE DIFFERENCE BETWEEN SIMPLY REACTING AND TRULY STRATEGIZING



2.

THE GAME OF SOUND

A new domain

WHAT IF AI LEARNED THE RULES OF SOUND?

In a game:
The AI sees visual frames and learns the game's physics.

In audio:
The AI 'sees' sound spectrograms and learns the 'rules of sound'.

3.

LEARNING BY PREDICTION

The core mechanic

HOW IT LEARNS

1. See the world

Convert audio into image-like spectrograms

2. Break it down

Divide the spectrogram into small time-based ‘tokens’

3. Predict

The model processes tokens to predict what comes next

THE REAL INNOVATION IS HOW THE MODEL LEARNS IF ITS PREDICTIONS ARE CORRECT

OLD WAY: RECONSTRUCTION

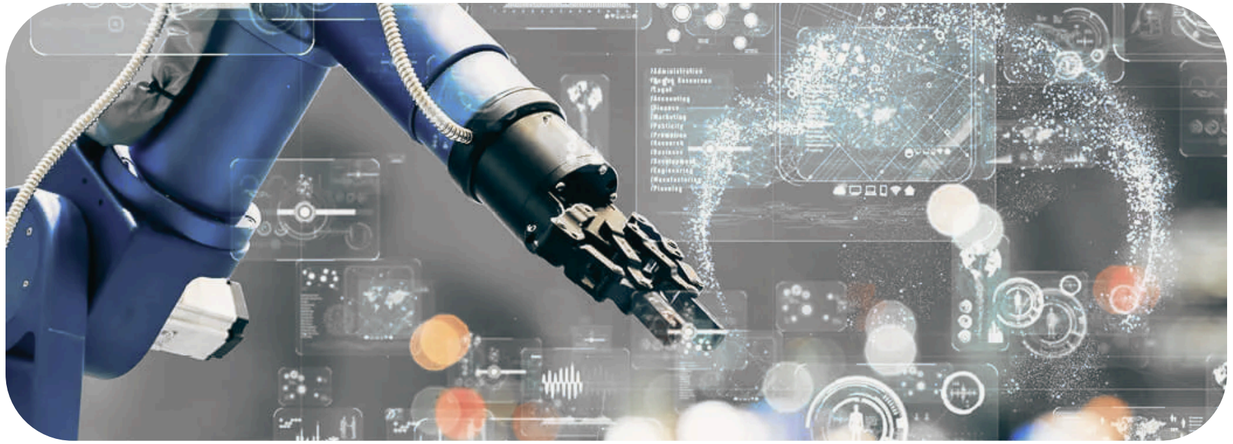
REQUIRED REDRAWING THE ENTIRE COMPLEX SPECTROGRAM FROM MEMORY

NEW WAY: CONTRASTIVE

LEARNS BY SYMPLE COMPARISON:
Is my prediction correct?

NO NEED TO RECONSTRUCT THE ENTIRE COMPLEX SPECTROGRAM

4.

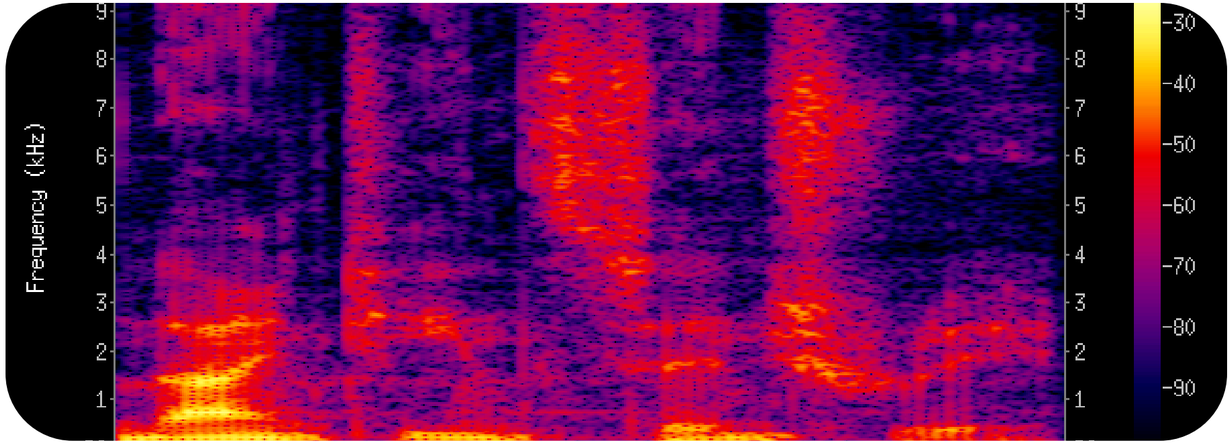


COMPOSING THE FUTURE

Potential applications

FUTURE POSSIBILITIES

- SYNTHETIZING COMPLETELY NEW AUDIO
- INTELLIGENTLY COMPLETING MISSING PARTS OF A SONG
- GENERATING NOVEL SOUNDS AND SOUNDTRACKS



5.

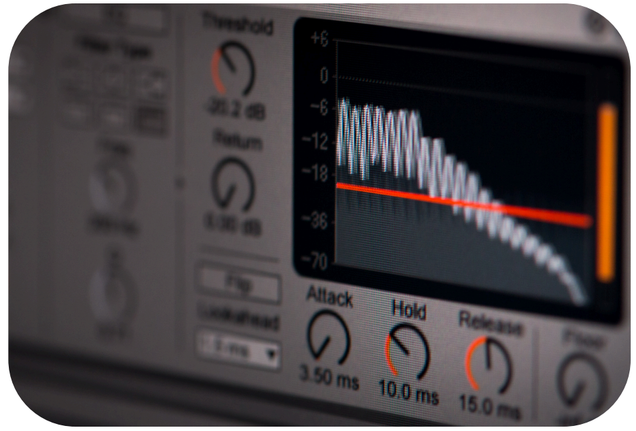
OUR PROJECT

CENTRAL IDEA

- ADAPT THE CONCEPT OF WORLD MODELS TO AUDIO SIGNAL ANALYSIS.
- GOAL: DEEP LEARNING MODEL TO LEARN REPRESENTATIONS & DYNAMICS OF AUDIO FROM SPECTROGRAMS.
- INSPIRATION: DREAMERV2 AND DREAMIN ARCHITECTURES.
- KEY ADAPTATION: REPLACE GAME IMAGES >> AUDIO SPECTROGRAMS.
- SPECTROGRAMS SPLIT INTO FIXED TEMPORAL WINDOWS (“TOKENS”)
- HYPOTHESIS: TOKENIZATION ENABLES SEQUENCE MODELING WITH ATTENTION ARCHITECTURES.

HOW DOES THIS WORK IN PRACTICE?

- MODEL LEARNS TRANSITIONS BETWEEN SPECTROGRAM TOKENS.
- CAPTURES TEMPORAL PROGRESSION AND RELATIONSHIPS AMONG SOUND ELEMENTS.
- OBJECTIVE: SHOW FEASIBILITY OF WORLD-MODEL ARCHITECTURE FOR AUDIO.
- FOCUS ON UNSUPERVISED LEARNING OF SIGNAL STRUCTURE, WITHOUT LARGE LABELED DATASETS.



EXPECTED RESULTS

- ARCHITECTURE TRAINED TO PREDICT THE NEXT SPECTROGRAM WINDOW FROM PREVIOUS ONES.
- TEST ABILITY TO GENERATE NEW SPECTROGRAM SEQUENCES:
  - COHERENT, TEMPORALLY STRUCTURED OUTPUTS >> EVIDENCE OF LEARNED DYNAMICS.
- LEARNED REPRESENTATION APPLICABLE TO:
  - AUDIO CLASSIFICATION (SPEECH, MUSIC, NOISE).
  - AUDIO GENERATION (VARIATIONS OF MELODIES).
  - SPEECH SYNTHESIS (LATENT STATES AS FOUNDATION).